ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Seminar for Applied Mathematics SAM

Efficient approximation of high-dimensional functions with deep neural networks

P. Cheridito and A. Jentzen and F. Rossmannek

Research Report No. 2019-64
December 2019

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

Efficient approximation of high-dimensional functions with deep neural networks

Patrick Cheridito* Arnulf Jentzen*† Florian Rossmannek*

December 12, 2019

Abstract

In this paper, we develop an approximation theory for deep neural networks that is based on the concept of a catalog network. Catalog networks are generalizations of standard neural networks in which the nonlinear activation functions can vary from layer to layer as long as they are chosen from a predefined catalog of continuous functions. As such, catalog networks constitute a rich family of continuous functions. We show that under appropriate conditions on the catalog, catalog networks can efficiently be approximated with neural networks and provide precise estimates on the number of parameters needed for a given approximation accuracy. We apply the theory of catalog networks to demonstrate that neural networks can overcome the curse of dimensionality in different high-dimensional approximation problems.

1 Introduction

It is well-known that neural networks with a single hidden layer can approximate any finite-dimensional continuous function on compact sets arbitrarily well if they are allowed to have sufficiently many hidden neurons; see e.g., Cybenko [3], Hornik et al. [10], Hornik [9], Leshno et al. [14], or Barron [1]. However, neural networks with more than one hidden layer typically show better performance in practical applications; see e.g., LeCun et al. [13] or Goodfellow et al. [5] and the references therein.

On the theoretical side, Eldan and Shamir [4] have provided an example of a simple continuous function that can be approximated much more efficiently with two hidden layers than with one. While this result holds for a large class of activation functions, Maiorov and Pinkus [15] have constructed a specific sigmoidal activation function that, in principle, allows to approximate every continuous function $f\colon [0,1]^d \to \mathbb{R}$ to any desired precision when used in a two-hidden-layers network with $3d$ neurons in the first and $6d+3$ neurons in the second hidden layer. Theoretically, this breaks the curse of dimensionality. But due to its complicated form, the activation function of Maiorov and Pinkus [15] is of little practical use. Moreover, it can be shown that their result only holds if the size of the network weights is allowed to grow faster than polynomially in the inverse of the approximation error; see e.g., Petersen and Voigtlaender [17]. So a better understanding of the approximation capacities of neural networks with commonly used activation functions is still of great interest. Mashkar [16] has shown that neural networks with multiple hidden layers and generalized sigmoidal activation functions are able to achieve the optimal rate of approximation for smooth and analytic functions. More recently, Petersen and Voigtlaender [17, 19] have derived the necessary complexity of ReLU networks needed for approximating classifier functions in L^p. L^2-approximation rates for different function classes are given in Bölcskei et al. [2] and Grohs et al. [8]. [6, 11, 12] have shown that solutions of various PDEs can be approximated with ReLU networks without the curse of dimensionality provided that the same is true for their coefficients and boundary conditions. In Schwab and Zech [18] neural network expression rates for generalized polynomial chaos expansions are given and it is shown that neural network can overcome the curse of dimensionality in the numerical approximation of solutions of certain parametric PDEs.

*Department of Mathematics, ETH Zürich

†Faculty of Mathematics and Computer Science, University of Münster

The purpose of this paper is to provide additional classes of high-dimensional functions that can efficiently be approximated with ReLU-like networks. To do that, we introduce the notion of a catalog network, which is a generalization of a standard feedforward network in which the nonlinear activation functions can vary from one layer to another as long as they are chosen from a given catalog of continuous functions. Particularly useful catalogs are catalogs of Lipschitz continuous functions, maximum functions, and products. We first study their approximability with neural networks. Then we show how the approximability of a catalog translates into the approximability of corresponding catalog networks. The theory of catalog networks can be used to construct different function classes that are approximable with ReLU-like neural networks with a number of parameters that is polynomial in the dimension and the inverse of the approximation accuracy.

The rest of the paper is organized as follows. In Section 2, we establish the notation, recall basic facts from [7, 12, 17] about the concatenation and parallelization of neural networks and derive two simple consequences that are needed later in the paper. In Section 3, we analyze the approximability of catalogs with neural networks and derive first consequences for the corresponding catalog networks. Section 4 is devoted to concrete examples of catalogs and a careful study of their approximability with neural networks. Section 5 contains the statement and proof of our main result, Theorem 5.3, which gives a precise estimate on the number of parameters needed to approximate a given catalog network to a desired accuracy with neural networks. In Section 6, we apply our main result to establish that different classes of high-dimensional functions are approximable with ReLU-like networks without the curse of dimensionality. Interestingly, in some cases, efficient approximation is possible with networks of constant depth as the dimension goes to infinity and the accuracy tends to zero, while in others our construction yields approximating networks with increasing depth.

2 Preliminaries

A neural network encodes a succession of affine and non-linear transformations. Denote $\mathbb{N} = \{1, 2, \dots\}$. Then the set of all neural network architectures is given by

$$\mathcal{N} = \cup_{D\in\mathbb{N}} \cup_{(l_0,\dots,l_D)\in\mathbb{N}^{D+1}} \times_{k=1}^{D} (\mathbb{R}^{l_k\times l_{k-1}} \times \mathbb{R}^{l_k}).$$

We denote the depth of a network architecture $\phi \in \mathcal{N}$ by $\mathcal{D}(\phi) = D$, the number of neurons in the k-th layer by $l_k^\phi = l_k$, $k \in \{0, \dots, D\}$, and the number of network parameters by $\mathcal{P}(\phi) = \sum_{k=1}^{D} l_k(l_{k-1} + 1)$. Moreover, if $\phi \in \mathcal{N}$ is given by $\phi = [(V_1, b_1), \dots, (V_D, b_D)]$, we denote by $\mathcal{A}_k^\phi \in C(\mathbb{R}^{l_{k-1}}, \mathbb{R}^{l_k})$, $k \in \{1, \dots, D\}$, the affine function $x \mapsto V_k x + b_k$. Let $a\colon \mathbb{R} \to \mathbb{R}$ be a continuous activation function. As usual, we extend it for every positive integer d, to a function from $\mathbb{R}^d$ to $\mathbb{R}^d$ mapping $(x_1, \dots, x_d)$ to $(a(x_1), \dots, a(x_d))$. The a-realization of $\phi \in \mathcal{N}$ is the function $\mathcal{R}_a^\phi \in C(\mathbb{R}^{l_0}, \mathbb{R}^{l_D})$ given by

$$\mathcal{R}_a^\phi = \mathcal{A}_D^\phi \circ a \circ \mathcal{A}_{D-1}^\phi \circ \cdots a \circ \mathcal{A}_1^\phi.$$

We recall that $\phi_1, \phi_2 \in \mathcal{N}$ can be composed such that the a-realization of the resulting network equals the concatenation $\mathcal{R}_a^{\phi_1} \circ \mathcal{R}_a^{\phi_2}$. This is done by combining the output layer of ϕ_2 with the input layer of ϕ_1. More precisely, if $\phi_1 = [(V_1, b_1), \dots, (V_D, b_D)]$ and $\phi_2 = [(W_1, c_1), \dots, (W_E, c_E)]$ satisfy $l_0^{\phi_1} = l_{\mathcal{D}(\phi_2)}^{\phi_2}$, then the concatenation $\phi_1 \circ \phi_2 \in \mathcal{N}$ is given by

$$\phi_1 \circ \phi_2 = [(W_1, c_1), \dots, (W_{E-1}, c_{E-1}), (V_1 W_E, V_1 c_E + b_1), (V_2, b_2), \dots, (V_D, b_D)].$$

The following result is straight-forward from the definition. A formal proof can be found in Grohs et al. [7].

Proposition 2.1. *The concatenation*

$$(\cdot) \circ (\cdot)\colon \{(\phi_1, \phi_2) \in \mathcal{N} \times \mathcal{N}\colon l_0^{\phi_1} = l_{\mathcal{D}(\phi_2)}^{\phi_2}\} \to \mathcal{N}$$

is associative and satisfies for all $\phi_1, \phi_2 \in \mathcal{N}$ with $l_0^{\phi_1} = l_{\mathcal{D}(\phi_2)}^{\phi_2}$

(i) $\mathcal{R}_a^{\phi_1 \circ \phi_2} = \mathcal{R}_a^{\phi_1} \circ \mathcal{R}_a^{\phi_2}$ *for all* $a \in C(\mathbb{R}, \mathbb{R})$,

(ii) $\mathcal{D}(\phi_1 \circ \phi_2) = \mathcal{D}(\phi_1) + \mathcal{D}(\phi_2) - 1$,

(iii) $l_k^{\phi_1 \circ \phi_2} = \begin{cases} l_k^{\phi_2} & \text{if } k \in \{0, \dots, \mathcal{D}(\phi_2) - 1\}, \\ l_{k+1-\mathcal{D}(\phi_2)}^{\phi_1} & \text{if } k \in \{\mathcal{D}(\phi_2), \dots, \mathcal{D}(\phi_1 \circ \phi_2)\}, \end{cases}$

(iv) $\mathcal{P}(\phi_1 \circ \phi_2) = \mathcal{P}(\phi_1) + \mathcal{P}(\phi_2) + l_1^{\phi_1} l_{\mathcal{D}(\phi_2)-1}^{\phi_2} - l_0^{\phi_1} l_1^{\phi_1} - l_{\mathcal{D}(\phi_2)}^{\phi_2} (l_{\mathcal{D}(\phi_2)-1}^{\phi_2} + 1)$,

(v) *if* $\mathcal{D}(\phi_1) = 1$, *then* $\mathcal{P}(\phi_1 \circ \phi_2) \leq \max\left\{1, l_{\mathcal{D}(\phi_1)}^{\phi_1} (l_{\mathcal{D}(\phi_2)}^{\phi_2})^{-1}\right\} \mathcal{P}(\phi_2)$,

(vi) *and if* $\mathcal{D}(\phi_2) = 1$, *then* $\mathcal{P}(\phi_1 \circ \phi_2) \leq \max\left\{1, (l_0^{\phi_1} + 1)^{-1} (l_0^{\phi_2} + 1)\right\} \mathcal{P}(\phi_1)$.

The next lemma is a direct consequence of the above and will be used later to estimate the number of parameters in our approximating networks.

Lemma 2.2. *Let* $a \in C(\mathbb{R}, \mathbb{R})$ *and* $\phi \in \mathcal{N}$. *Suppose that* $\psi_1, \psi_2 \in \mathcal{N}$ *satisfy* $\mathcal{D}(\psi_1) = \mathcal{D}(\psi_2) = 2$, $l_0^{\psi_1} = l_2^{\psi_1} = l_{\mathcal{D}(\phi)}^{\phi}$, *and* $l_0^{\psi_2} = l_2^{\psi_2} = l_0^{\phi}$. *Abbreviate* $D = \mathcal{D}(\phi)$. *Then*

$$\mathcal{P}(\psi_1 \circ \phi \circ \psi_2) = \begin{cases} \mathcal{P}(\phi) + l_1^{\psi_2}(l_0^{\phi} + 1) + l_1^{\psi_1}(l_D^{\phi} + 1) + l_1^{\psi_1} l_1^{\psi_2} - l_0^{\phi} l_D^{\phi} & \text{if } D = 1, \\ \mathcal{P}(\phi) + l_1^{\psi_2}(l_0^{\phi} + 1) + l_1^{\psi_1}(l_D^{\phi} + 1) + l_1^{\phi}(l_1^{\psi_2} - l_0^{\phi}) + l_{D-1}^{\phi}(l_1^{\psi_1} - l_D^{\phi}) & \text{if } D \geq 2. \end{cases}$$

Proof. Let $k \in \mathbb{N}$ be given by $k = l_1^{\psi_2}$ if $D = 1$ and $k = l_{D-1}^{\phi}$ if $D \geq 2$. By Proposition 2.1 and the fact that $\mathcal{P}(\psi_2) = l_1^{\psi_2}(l_0^{\psi_2} + 1) + l_2^{\psi_2}(l_1^{\psi_2} + 1)$, we have

$$\mathcal{P}(\phi \circ \psi_2) = \mathcal{P}(\phi) + l_1^{\psi_2}(l_0^{\phi} + 1) + l_1^{\phi}(l_1^{\psi_2} - l_0^{\phi})$$

and also $l_D^{\phi \circ \psi_2} = k$. Thus, by applying Proposition 2.1 once more and observing $l_{D+1}^{\phi \circ \psi_2} = l_D^{\phi} = l_2^{\psi_1}$, we obtain

$$\begin{aligned} \mathcal{P}(\psi_1 \circ \phi \circ \psi_2) &= \mathcal{P}(\phi \circ \psi_2) + l_1^{\psi_1}(l_D^{\phi} + 1) + k(l_1^{\psi_1} - l_D^{\phi}) \\ &= \mathcal{P}(\phi) + l_1^{\psi_2}(l_0^{\phi} + 1) + l_1^{\psi_1}(l_D^{\phi} + 1) + l_1^{\phi}(l_1^{\psi_2} - l_0^{\phi}) + k(l_1^{\psi_1} - l_D^{\phi}), \end{aligned}$$

which completes the proof of Lemma 2.2. □

Another operation on neural networks that we will need is parallelization. In the case that $\phi_1 = [(V_1, b_1), \dots, (V_D, b_D)]$ and $\phi_2 = [(W_1, c_1), \dots, (W_D, c_D)]$ have the same depth, then this can be achieved by constructing block matrices in each layer by

$$p(\phi_1, \phi_2) = \left[\left(\begin{bmatrix} V_1 & 0 \\ 0 & W_1 \end{bmatrix}, \begin{bmatrix} b_1 \\ c_1 \end{bmatrix} \right), \dots, \left(\begin{bmatrix} V_D & 0 \\ 0 & W_D \end{bmatrix}, \begin{bmatrix} b_D \\ c_D \end{bmatrix} \right) \right].$$

Clearly, we can then define the parallelization of arbitrarily many neural networks $\phi_1, \dots, \phi_n \in \mathcal{N}$, $n \in \mathbb{N}$, with the same depth by iteration

$$p(\phi_1, \dots, \phi_n) = p(\phi_1, p(\phi_2, p(\dots, \phi_n))) \dots).$$

Statements (i)–(ii) in the next proposition follow immediately from the definition. For (iii)–(iv), we refer to Grohs et al. [7].

Proposition 2.3. *The parallelization*

$$p \colon \cup_{n \in \mathbb{N}} \{(\phi_1, \dots, \phi_n) \in \mathcal{N}^n \colon \mathcal{D}(\phi_1) = \dots = \mathcal{D}(\phi_n)\} \to \mathcal{N}$$

satisfies for all $n \in \mathbb{N}$ *and* $\phi_1, \dots, \phi_n \in \mathcal{N}$ *with the same depth:*

(i) $\mathcal{R}_a^{p(\phi_1,\dots,\phi_n)}((x_1,\dots,x_n)) = (\mathcal{R}_a^{\phi_1}(x_1),\dots,\mathcal{R}_a^{\phi_n}(x_n))$ *for all* $x_1 \in \mathbb{R}^{l_0^{\phi_1}},\dots,x_n \in \mathbb{R}^{l_0^{\phi_n}}$ *and each* $a \in C(\mathbb{R},\mathbb{R})$,

(ii) $l_k^{p(\phi_1,\dots,\phi_n)} = \sum_{j=1}^n l_k^{\phi_j}$ *for all* $k \in \{0,\dots,\mathcal{D}(\phi_1)\}$,

(iii) $\mathcal{P}(p(\phi_1,\dots,\phi_n)) \leq \frac{1}{2}\big[\sum_{j=1}^n \mathcal{P}(\phi_j)\big]^2$,

(iv) $\mathcal{P}(p(\phi_1,\dots,\phi_n)) \leq n^2\mathcal{P}(\phi_1)$ *whenever* $l_k^{\phi_i} = l_k^{\phi_j}$ *for each* $k \in \{0,\dots,\mathcal{D}(\phi_1)\}$ *and all* $i,j \in \{1,\dots,n\}$.

The above construction needs all the neural networks to have the same depth. If this is not the case, then we can still parallelize neural networks but only for a special class of activation functions.

Definition 2.4. *We say a function* $a \in C(\mathbb{R},\mathbb{R})$ *fulfills the c-identity requirement for a number* $c \geq 2$ *if there exists* $I \in \mathcal{N}$ *such that* $\mathcal{D}(I) = 2$, $l_1^I \leq c$ *and* $\mathcal{R}_a^I = \mathrm{id}_{\mathbb{R}}$.

Note that if I satisfies $\mathcal{R}_a^I = \mathrm{id}_{\mathbb{R}}$, one can also realize the identity function $\mathrm{id}_{\mathbb{R}^d}$ in d dimensions for any $d \in \mathbb{N}$, using d-fold parallelization $I_d = p(I,\dots,I)$. The c-identity requirement is crucial for our purposes. The main example we have in mind is the ReLU activation $\mathrm{ReLU}\colon \mathbb{R} \to \mathbb{R},\ x \mapsto \max\{x,0\}$, which satisfies the 2-identity requirement with $I = [([1\ \ -1]^T, [0\ 0]^T), ([1\ \ -1], 0)]$. However, it is easy to see that generalized ReLU functions of the form

$$a(x) = \begin{cases} rx & \text{if } x \geq 0 \\ sx & \text{if } x < 0 \end{cases}$$

for $(r,s) \in \mathbb{R}_+^2 \setminus \{0\}$, also satisfy the 2-identity requirement.

Using the identity requirement, we can extend the notion of parallelization. Namely, if $\phi_1,\dots,\phi_n \in \mathcal{N}$, $n \in \mathbb{N}$, do not have the same depth, then we simply concatenate the shorter networks with the identity I_d until they all have the same depth. Then one can apply the original parallelization. Of course, the parameter count changes compared to simple parallelization. The following result follows from Grohs et al. [7, Corollary 2.24].

Proposition 2.5. *Assume* $a \in C(\mathbb{R},\mathbb{R})$ *fulfills the c-identity requirement for a number* $c \geq 2$ *with* $I \in \mathcal{N}$. *Then the extended parallelization* $p_I\colon \cup_{n\in\mathbb{N}} \mathcal{N}^n \to \mathcal{N}$ *satisfies*

$$\mathcal{P}(p_I(\phi_1,\dots,\phi_n)) \leq \frac{1}{2}\Big[\sum_{j=1}^n c\mathcal{P}(\phi_j) + cl_{\mathcal{D}(\phi_j)}^{\phi_j}(cl_{\mathcal{D}(\phi_j)}^{\phi_j}+1)\max_{i\in\{1,\dots,n\}}\mathcal{D}(\phi_i)\Big]^2$$

for all $n \in \mathbb{N}$ *and* $\phi_1,\dots,\phi_n \in \mathcal{N}$.

We finish this section with the following consequence of Lemma 2.2.

Corollary 2.6. *Assume* $a \in C(\mathbb{R},\mathbb{R})$ *satisfies the c-identity requirement for a number* $c \geq 2$ *with* $I \in \mathcal{N}$. *Let* $\phi \in \mathcal{N}$ *and denote the d-fold parallelization of* I *by* I_d. *Abbreviate* $m = l_0^\phi$, $n = l_{\mathcal{D}(\phi)}^\phi$, *and* $k = \max\{m,n\}$. *Then*

$$\mathcal{P}(I_n \circ \phi \circ I_m) \leq \mathcal{P}(\phi)ck + 3c^2k^2.$$

Proof. Abbreviate $D = \mathcal{D}(\phi)$. Lemma 2.2 yields

$$\mathcal{P}(I_n \circ \phi \circ I_m) = \begin{cases} \mathcal{P}(\phi) + l_1^{I_m}(m+1) + l_1^{I_n}(n+1) + l_1^{I_m}l_1^{I_n} - mn & \text{if } D = 1, \\ \mathcal{P}(\phi) + l_1^{I_m}(m+1) + l_1^{I_n}(n+1) + l_1^\phi(l_1^{I_m} - m) + l_{D-1}^\phi(l_1^{I_n} - n) & \text{if } D \geq 2. \end{cases}$$

Note that the hypothesis $l_1^I \leq c$ and item (ii) in Proposition 2.3 ensure that $l_1^{I_m}$ and $l_1^{I_n}$ are both at most ck. This and the fact that $l_1^\phi + l_{D-1}^\phi \leq \mathcal{P}(\phi)$ imply

$$\begin{aligned} \mathcal{P}(I_n \circ \phi \circ I_m) &\leq \begin{cases} \mathcal{P}(\phi) + 2ck(k+1) + c^2k^2 - k & \text{if } D = 1, \\ \mathcal{P}(\phi) + 2ck(k+1) + (l_1^\phi + l_{D-1}^\phi)(ck-1) & \text{if } D \geq 2 \end{cases} \\ &\leq \begin{cases} \mathcal{P}(\phi) + 3c^2k^2 & \text{if } D = 1, \\ \mathcal{P}(\phi)ck + 3c^2k^2 & \text{if } D \geq 2, \end{cases} \end{aligned}$$

where last inequality holds because $c \geq 2$. ∎

3 Catalog Networks

In this section, we generalize the concept of a neural network by allowing the activation functions to change from layer to layer. But they have to be chosen from a predefined catalog $\mathcal{F} \subseteq \cup_{m,n\in\mathbb{N}} C(\mathbb{R}^m, \mathbb{R}^n)$. The following notation will be useful: We denote the dimension of the domain of a function $f \in \cup_{m,n\in\mathbb{N}} C(\mathbb{R}^m, \mathbb{R}^n)$ by $d_0(f)$ and the dimension of its target space by $d_1(f)$, so that $f \in C(\mathbb{R}^{d_0(f)}, \mathbb{R}^{d_1(f)})$. Now, consider a catalog $\mathcal{F}$ as well as $D \in \mathbb{N}$ and $l_0, \dots, l_{2D} \in \mathbb{N}$. Then we define $\mathcal{C}_{\mathcal{F}}^{l_0,\dots,l_{2D}}$ to be the set

$$\mathcal{C}_{\mathcal{F}}^{l_0,\dots,l_{2D}} = \bigtimes_{n=1}^{D} \Bigg[\mathbb{R}^{l_{2n-1}\times l_{2n-2}} \times \mathbb{R}^{l_{2n-1}} \times \bigcup_{k\in\mathbb{N}} \Big\{ (f_1,\dots,f_k) \in \mathcal{F}^k \colon \Big[\text{for all } i \in \{0,1\} \colon \sum_{j=1}^{k} d_i(f_j) = l_{2n-1+i} \Big] \Big\} \Bigg].$$

As in the definition of neural network architectures, there are affine transformations encoded in the first two components of the inner cartesian product. The last term consists of a tuple of continuous functions $f_1, \dots, f_k$, which are applied in a parallelized way in place of the activation function after each affine transformation. We define the set of all catalog networks corresponding to $\mathcal{F}$ by

$$\mathcal{C}_{\mathcal{F}} = \bigcup_{D\in\mathbb{N}} \bigcup_{l_0,\dots,l_{2D}\in\mathbb{N}} \mathcal{C}_{\mathcal{F}}^{l_0,\dots,l_{2D}}.$$

We define the depth of a catalog network $\xi \in \mathcal{C}_{\mathcal{F}}^{l_0,\dots,l_{2D}}$ as $\mathcal{D}_C(\xi) = D$. Its input dimension is $\mathcal{I}_C(\xi) = l_0$, its output dimension $\mathcal{O}_C(\xi) = l_{2D}$, and its maximal width $\mathcal{W}_C(\xi) = \max\{l_0, \dots, l_{2D}\}$. Next, we discuss the realization of a catalog network. Suppose $\xi \in \mathcal{C}_{\mathcal{F}}^{l_0,\dots,l_{2D}}$ is given by $\xi = [(V_1, b_1, (f_{1,1}, \dots, f_{1,k_1})), \dots, (V_D, b_D, (f_{D,1}, \dots, f_{D,k_D}))]$. Then we let $\mathcal{A}_C^{\xi,n} \in C(\mathbb{R}^{l_{2n-2}}, \mathbb{R}^{l_{2n-1}})$, $n \in \{1, \dots, D\}$, be the affine function $x \mapsto V_n x + b_n$. By $\mathcal{G}_C^{\xi,n} \in C(\mathbb{R}^{l_{2n-1}}, \mathbb{R}^{l_{2n}})$, $n \in \{1, \dots, D\}$, we denote the function mapping $x \in \mathbb{R}^{l_{2n-1}}$ to

$$\mathcal{G}_C^{\xi,n}(x) = \Big(f_{n,1}\big(x_1, \dots, x_{d_0(f_{n,1})}\big), f_{n,2}\big(x_{d_0(f_{n,1})+1}, \dots, x_{d_0(f_{n,1})+d_0(f_{n,2})}\big), \dots f_{n,k_n}\big(x_{d_0(f_{n,1})+\dots+d_0(f_{n,k_n-1})+1}, \dots, x_{d_0(f_{n,1})+\dots+d_0(f_{n,k_n})}\big)\Big),$$

that is, we apply $f_{n,1}$ to the first $d_0(f_{n,1})$ entries of x, $f_{n,2}$ to the next $d_0(f_{n,2})$ entries, and so on. This is well-defined due to the condition $d_i(f_1) + \dots + d_i(f_k) = l_{2n-1+i}$, $i \in \{0,1\}$, posed in the definition of $\mathcal{C}_{\mathcal{F}}^{l_0,\dots,l_{2D}}$. The overall realization function $\mathcal{R}_C^{\xi} \in C(\mathbb{R}^{l_0}, \mathbb{R}^{l_{2D}})$ generated by the catalog network ξ is

$$\mathcal{R}_C^{\xi} = \mathcal{G}_C^{\xi,D} \circ \mathcal{A}_C^{\xi,D} \circ \dots \circ \mathcal{G}_C^{\xi,1} \circ \mathcal{A}_C^{\xi,1}.$$

Our goal is to show that catalog networks can efficiently be approximated with neural networks with respect to some weight function, by which we mean any function $w \colon [0,\infty) \to [0,\infty)$.

Definition 3.1. *We say a weight function w has order of growth at most $(s_1, s_2) \in [1,\infty) \times [0,\infty)$ if*

$$w(x) \le s_1 r^{s_2} w(r \max\{x, 1\})$$

for all $x \in [0,\infty)$ and $r \in [1,\infty)$.

Useful weight functions are constants and functions of the form $(1 + x^q)^{-1}$ or $(\max\{1, x^q\})^{-1}$ for some $q \in (0,\infty)$. Constant weight functions clearly have order of growth at most $(1,0)$. The order of growth of $(1 + x^q)^{-1}$ and $(\max\{1, x^q\})^{-1}$ follows from Lemma 3.2 below. The order of growth is a general concept applicable to different types of weight functions. The inequality in Definition 3.1 is exactly what is needed in the proof of our main result. Note that if a weight function w has an order of growth and satisfies $w(x) = 0$ for some $x \in [1,\infty)$, then $w(y) = 0$ for all $y \in [0, x]$. In particular, indicator functions of bounded intervals cannot have an order of growth. We address this issue by introducing approximation sets in Definition 3.3.

Lemma 3.2. *Let $\delta \in (0,\infty)$ and consider a non-decreasing function $f\colon [0,\infty) \to (0,\infty)$. Moreover, let $p\colon [0,\infty) \to [0,\infty)$ be of the form $x \mapsto \sum_{k=0}^{q} a_k x^{b_k}$, where $q \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}$ and $a_0, b_0, \dots, a_q, b_q \in [0,\infty)$. Then the weight function w given by $w(x) = f(x)(\max\{p(x),\delta\})^{-1}$ has order of growth at most $(\max\{\frac{p(1)}{\delta}, 1\}, \max\{b_0,\dots,b_q\})$.*

Proof. Abbreviate $s = \max\{b_0,\dots,b_q\}$ and note that p having non-negative coefficients $a_0,\dots,a_q$ implies for all $x \in [0,\infty)$, $r \in [1,\infty)$ that $p(rx) \le r^s p(x)$. This and the hypothesis that f is non-decreasing yield for all $x \in [0,\infty)$, $r \in [1,\infty)$

$$w(x) \le \frac{f(rx)}{\max\{p(x),\delta\}} \le \frac{f(rx)r^s}{\max\{p(rx),\delta\}} = r^s w(rx).$$

Next, we use the hypothesis that f is non-decreasing again to find for all $x \in [0,1)$

$$w(x) \le \frac{f(1)}{\max\{p(x),\delta\}} \le \frac{\max\{p(1),\delta\}}{\delta} w(1).$$

Combining the previous two calculations yields

$$\frac{\delta}{\max\{p(1),\delta\}} w(x) \le w(\max\{x,1\}) \le r^s w(\max\{x,1\}),$$

which concludes Lemma 3.2. □

We are primarily interested in catalogs of functions that are well approximable with neural networks. In addition, we require the approximations to be Lipschitz continuous with a Lipschitz constant independent of the accuracy. Let us make this precise. We denote by $\|\cdot\|$ the Euclidean norm.

Definition 3.3. *Let $a \in C(\mathbb{R},\mathbb{R})$ and consider a weight function w. Fix $L \in [0,\infty)$ and $\varepsilon \in (0,1]$. Given a function $f \in \cup_{m,n\in\mathbb{N}} C(\mathbb{R}^m,\mathbb{R}^n)$ and any set $B \subseteq \mathbb{R}^{d_0(f)}$, we define the cost to approximate f with accuracy ε and weight w with L-Lipschitz neural networks on the set B as*

$$\mathrm{Cost}_{a,w}(f,B,L,\varepsilon) = \inf\left\{ \mathcal{P}(\phi) \in \mathbb{N}\colon \left[\begin{array}{c} \phi \in \mathcal{N} \text{ with } \mathcal{R}_a^\phi \in C(\mathbb{R}^{d_0(f)},\mathbb{R}^{d_1(f)}) \\ \text{s.t. } \mathcal{R}_a^\phi \text{ is } L\text{-Lipschitz on } \mathbb{R}^{d_0(f)} \text{ and} \\ \sup_{x\in B} w(\|x\|)\|f(x) - \mathcal{R}_a^\phi(x)\| \le \varepsilon \end{array} \right] \right\},$$

where we use the usual convention $\inf(\emptyset) = \infty$.

The next definition embodies the class of catalogs we will use for our catalog networks.

Definition 3.4. *Let $a \in C(\mathbb{R},\mathbb{R})$, $\kappa = (\kappa_0,\kappa_1,\kappa_2,\kappa_3) \in [1,\infty)^2 \times [0,\infty)^2$, $\varepsilon \in (0,1]$, and suppose w is a weight function that never vanishes. Consider a family of sets $B = (B_f)_{f\in\mathcal{F}}$ such that $B_f \subseteq \mathbb{R}^{d_0(f)}$ contains the origin for all $f \in \mathcal{F}$, and let $L = (L_f)_{f\in\mathcal{F}} \subseteq [0,\infty)$ be a collection of Lipschitz constants. Then we call a subset $\mathcal{F} \subseteq \cup_{m,n\in\mathbb{N}} C(\mathbb{R}^m,\mathbb{R}^n)$ an $[a,w,B,L,\varepsilon,\kappa]$-approximable catalog if $\sup_{f\in\mathcal{F}} \|f(0)\| \le \kappa_0$ and*

$$\mathrm{Cost}_{a,w}(f,B_f,L_f,\delta) \le \kappa_1 |\max\{d_0(f),d_1(f)\}|^{\kappa_2} \delta^{-\kappa_3}$$

for all $f \in \mathcal{F}$ and $\delta \in (0,\varepsilon]$.

Note that if $\mathcal{F}$ is $[a,w,B,L,\varepsilon,\kappa]$-approximable, then every $f \in \mathcal{F}$ must be L_f-Lipschitz continuous on the set B_f. Indeed, the definition implies that for all $\delta \in (0,\varepsilon]$ there exists a $\phi_\delta \in \mathcal{N}$ such that $w(\|x\|)\|f(x) - \mathcal{R}_a^{\phi_\delta}(x)\| \le \delta$ and $\|\mathcal{R}_a^{\phi_\delta}(x) - \mathcal{R}_a^{\phi_\delta}(y)\| \le L_f\|x-y\|$ for all $x,y \in B_f$. So, one obtains from the triangle inequality that

$$\|f(x) - f(y)\| \le \frac{\delta}{w(\|x\|)} + L_f\|x-y\| + \frac{\delta}{w(\|y\|)}$$

for all $x, y \in B_f$ and $\delta > 0$.

Coming back to catalog networks, we would like to deduce a Lipschitz property for the generalized activation functions $\mathcal{G}_C^{\xi,n}$ in a catalog network $\xi \in \mathcal{C}_{\mathcal{F}}$ corresponding to an approximable catalog $\mathcal{F}$. To do so, we need two more definitions. Let $\mathcal{F}$ be a catalog approximable on sets $B = (B_f)_{f\in\mathcal{F}}$ with Lipschitz constants $L = (L_f)_{f\in\mathcal{F}} \subseteq [0,\infty)$. For a catalog network $\xi \in \mathcal{C}_{\mathcal{F}}^{l_0,\dots,l_{2D}}$ given by $\xi = [(V_1, b_1, (f_{1,1},\dots,f_{1,k_1})),\dots,(V_D, b_D, (f_{D,1},\dots,f_{D,k_D}))]$, we set

$$\mathbb{B}_B^{\xi,n} := \bigtimes_{j=1}^{k_n} B_{f_{n,j}} \subseteq \bigtimes_{j=1}^{k_n} \mathbb{R}^{d_0(f_{n,j})} = \mathbb{R}^{l_{2n-1}}$$

and

$$L^{\xi,n} := \max_{j\in\{1,\dots,k_n\}} L_{f_{n,j}}.$$

$\mathbb{B}_B^{\xi,n}$ is the set on which all the functions used in the n-th layer can be approximated. Moreover, the following holds.

Lemma 3.5. *Let $\xi \in \mathcal{C}_{\mathcal{F}}$ be a catalog network based on an $[a, w, B, L, \varepsilon, \kappa]$-approximable catalog $\mathcal{F}$. Then*

$$\|\mathcal{G}_C^{\xi,n}(x) - \mathcal{G}_C^{\xi,n}(y)\| \le L^{\xi,n}\|x - y\|$$

for all $n \in \{1,\dots,\mathcal{D}_C(\xi)\}$ and $x, y \in \mathbb{B}_B^{\xi,n}$.

Proof. Assume that $\xi = [(V_1, b_1, (f_{1,1},\dots,f_{1,k_1})),\dots,(V_D, b_D, (f_{D,1},\dots,f_{D,k_D}))]$. The discussion after Definition 3.4 showed that every $f \in \mathcal{F}$ is L_f-Lipschitz continuous on the set B_f. Given $n \in \{1,\dots,D\}$, $j \in \{1,\dots,k_n\}$, and $x \in \mathbb{R}^{l_{2n-1}}$, denote by $x_{(n,j)} \in \mathbb{R}^{d_0(f_{n,j})}$ the vector

$$x_{(n,j)} = \left(x_{d_0(f_{n,1})+\cdots+d_0(f_{n,j-1})+1},\dots,x_{d_0(f_{n,1})+\cdots+d_0(f_{n,j})}\right).$$

Then one has for all $n \in \{1,\dots,D\}$ and $x, y \in \mathbb{B}_B^{\xi,n}$,

$$\begin{aligned}\|\mathcal{G}_C^{\xi,n}(x) - \mathcal{G}_C^{\xi,n}(y)\|^2 &= \sum_{j=1}^{k_n} \|f_{n,j}(x_{(n,j)}) - f_{n,j}(y_{(n,j)})\|^2 \\ &\le \sum_{j=1}^{k_n} |L_{f_{n,j}}|^2 \|x_{(n,j)} - y_{(n,j)}\|^2 \le |L^{\xi,n}|^2\|x-y\|^2.\end{aligned} \tag{3.1}$$

□

4 Examples of approximable catalogs

In this section, we provide different examples of approximable catalogs that will be used in Section 6 to show that certain high-dimensional functions are approximable without the curse of dimensionality. We focus on one-dimensional Lipschitz functions, the maximum function in arbitrary dimension, and the product function in two dimensions.

First, consider a K-Lipschitz function $f\colon \mathbb{R} \to \mathbb{R}$ for some $K \in (0,\infty)$ such that $|f(0)| \le K$. For any given $r \in (0,\infty)$, f can be approximated on $[-r, r]$ with a piece-wise linear function supported on $N+1$ equidistributed points with accuracy Kr/N. Such a piece-wise linear function can be realized with a ReLU network ϕ_N with one hidden layer and N hidden neurons. This results in $\mathcal{P}(\phi_N) = 3N + 1$, and it follows that

$$\mathrm{Cost}_{\mathrm{ReLU},w_0}(f, [-r, r], K, \varepsilon) \le \mathcal{P}(\phi_{\lceil Kr\varepsilon^{-1}\rceil}) \le 3Kr\varepsilon^{-1} + 4,$$

where w_0 is the weight function that is constantly equal to 1. Alternatively, one can approximate f on the entire real line with respect to a weight function of the form $w_q(x) = (1+x^q)^{-1}$ for some $q \in (1,\infty)$. Then

$$\mathrm{Cost}_{\mathrm{ReLU},w_q}(f, \mathbb{R}, K, \varepsilon) \le (6K)^{\frac{q}{q-1}}\varepsilon^{-\frac{q}{q-1}} + 10;$$

see Hutzenthaler et al. [11, Corollary 3.13].

Next, we turn to the maximum functions $\max_d\colon \mathbb{R}^d \to \mathbb{R},\ x \mapsto \max\{x_1, \dots, x_d\}$, $d \in \mathbb{N}$. These can be represented exactly by ReLU networks. $\max_1$ is simply the identity and $\max_2$ is the ReLU-realization of

$$\phi_2 = \left[\left(\begin{bmatrix} 1 & -1 \\ 0 & 1 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \right), \left(\begin{bmatrix} 1 & 1 & -1 \end{bmatrix}, 0 \right) \right].$$

If $I \in \mathcal{N}$ is taken such that ReLU fulfills the 2-identity requirement with I and we define $I_d = p(I, \dots, I)$, $d \in \mathbb{N}$, then it can easily be shown by induction that $\max_d$, $d \in \mathbb{N}_{\geq 3}$, is the ReLU-realization of $\phi_d = \phi_{d-1} \circ p(\phi_2, I_{d-2})$ and $\mathcal{P}(\phi_d) = \frac{1}{3}(4d^3 + 3d^2 - 4d + 3) \leq 2d^3$. In other words, for all $d \in \mathbb{N}$ and any weight function w

$$\mathrm{Cost}_{\mathrm{ReLU},w}(\max_d, \mathbb{R}^d, 1, \varepsilon) \leq 2d^3 = 2|d_0(\max_d)|^3.$$

Finally, we study the approximability of the product function. To do that, we first consider the square function $\mathrm{sq}\colon \mathbb{R} \to \mathbb{R},\ x \mapsto x^2$. It has been shown by different authors that the square function can be approximated with accuracy $\varepsilon > 0$ on the unit interval by a ReLU network ϕ_ε satisfying $\mathcal{P}(\phi_\varepsilon) = \mathcal{O}(\log_2(\varepsilon^{-1}))$; see [7, 8, 18, 20]. This relies on realizing linear combinations of iterations of the sawtooth function with a neural network and establishing exponentially fast convergence in the number of iterations. In particular, the approximator $\mathcal{R}^{\phi}_{\mathrm{ReLU}}$ is 2-Lipschitz. A precise estimate of the number of parameters required is given in Grohs et al. [7, Proposition 3.3]. In our language it can be stated as

$$\mathrm{Cost}_{\mathrm{ReLU},w_0}(\mathrm{sq}, [0,1], 2, \varepsilon) \leq \mathcal{P}(\phi_\varepsilon) \leq \max\{13, 10\log_2(\varepsilon^{-1}) - 7\}.$$

Moreover, the neural network ϕ_ε achieving this Cost also satisfies $\mathcal{R}^{\phi_\varepsilon}_{\mathrm{ReLU}} = \mathrm{ReLU}$ on $\mathbb{R}\backslash[0,1]$. By a mirroring argument, we can then obtain an approximation of the square function on the interval $[-r, r]$ for any $r \in (0, \infty)$.

Lemma 4.1. *For all $r \in (0,\infty)$ and $\varepsilon \in (0,1]$, there exists a neural network $\psi_{r,\varepsilon} \in \mathcal{N}$ such that $\mathcal{R}^{\psi_{r,\varepsilon}}_{\mathrm{ReLU}} \in C(\mathbb{R},\mathbb{R})$ is $2r$-Lipschitz continuous, $\sup_{x\in[-r,r]} |\mathcal{R}^{\psi_{r,\varepsilon}}_{\mathrm{ReLU}}(x) - x^2| \leq \varepsilon$, $\mathcal{R}^{\psi_{r,\varepsilon}}_{\mathrm{ReLU}}(x) = r|x|$ for all $x \in \mathbb{R}\backslash[-r,r]$, and*

$$\mathcal{P}(\psi_{r,\varepsilon}) \leq \max\{52, 80\log_2(r) + 40\log_2(\varepsilon^{-1}) - 28\}.$$

Proof. Take $\phi_{r,1} \in \mathcal{N}$ of depth 1 with $\mathcal{R}^{\phi_{r,1}}_{\mathrm{ReLU}} \in C(\mathbb{R}^2,\mathbb{R})$ given by $(x,y) \mapsto r^2(x+y)$ and take $\phi_{r,2} \in \mathcal{N}$ of depth 1 with $\mathcal{R}^{\phi_{r,2}}_{\mathrm{ReLU}} \in C(\mathbb{R},\mathbb{R}^2)$ given by $x \mapsto (\frac{x}{r}, -\frac{x}{r})$. If $(\phi_\varepsilon)_{\varepsilon\in(0,1]} \subseteq \mathcal{N}$ denote the approximators of the square function on the unit interval from above, then $\psi_{r,\varepsilon} = \phi_{r,1} \circ p(\phi_{r^{-2}\varepsilon}, \phi_{r^{-2}\varepsilon}) \circ \phi_{r,2}$ approximates the square function on $[-r,r]$ with accuracy ε. To see this, note that $\mathcal{R}^{\psi_{r,\varepsilon}}_{\mathrm{ReLU}}(x) = r^2\mathcal{R}^{\phi_{r^{-2}\varepsilon}}_{\mathrm{ReLU}}(\frac{|x|}{r})$ for all $x \in \mathbb{R}$ since $\phi_{r^{-2}\varepsilon} = \mathrm{ReLU}$ on $\mathbb{R}\backslash[0,1]$. Moreover, this also implies $\mathcal{R}^{\psi_{r,\varepsilon}}_{\mathrm{ReLU}}(x) = r|x|$ for all $x \in \mathbb{R}\backslash[-r,r]$ as well as the Lipschitz continuity. Lastly, Proposition 2.1 (items (v) and (vi)) and Proposition 2.3 (item (iv)) assure $\mathcal{P}(\psi_{r,\varepsilon}) \leq 4\mathcal{P}(\phi_{r^{-2}\varepsilon})$. This finishes the proof of Lemma 4.1. □

More concisely, the statement of Lemma 4.1 can be written as

$$\mathrm{Cost}_{\mathrm{ReLU},w_0}(\mathrm{sq}, [-r,r], 2r, \varepsilon) \leq \max\{52, 80\log_2(r) + 40\log_2(\varepsilon^{-1}) - 28\}.$$

We can now estimate the approximation rate of the product function $\mathrm{pr}\colon \mathbb{R}^2 \to \mathbb{R},\ (x,y) \mapsto xy$ using the identity $xy = \frac{1}{2}((x+y)^2 - x^2 - y^2)$. This trick has already been used before; see [7, 8, 18, 20]. But let us still provide a proof of the next proposition, since most results in the existing literature are stated slightly differently or do not specify the Lipschitz constant.

Proposition 4.2. *For all $r \in (0,\infty)$ and $\varepsilon \in (0,1]$, one has*

$$\mathrm{Cost}_{\mathrm{ReLU},w_0}(\mathrm{pr}, [-r,r]^2, \sqrt{32}r, \varepsilon) \leq \max\{468, 679 + 720\log_2(r) + 360\log_2(\varepsilon^{-1})\}.$$

Proof. Pick $\psi_1 \in \mathcal{N}$ of depth 1 with $\mathcal{R}^{\psi_1}_{\mathrm{ReLU}} \in C(\mathbb{R}^3, \mathbb{R})$ given by $(x, y, z) \mapsto \frac{1}{2}(x - y - z)$ and pick $\psi_2 \in \mathcal{N}$ of depth 1 with $\mathcal{R}^{\psi_2}_{\mathrm{ReLU}} \in C(\mathbb{R}^2, \mathbb{R}^3)$ given by $(x, y) \mapsto (x + y, x, y)$. If $(\psi_{r,\varepsilon})_{r\in(0,\infty),\varepsilon\in(0,1]} \subseteq \mathcal{N}$ denote the approximators of the square function on the interval $[-r, r]$ from Lemma 4.1, then $\chi_{r,\varepsilon} = \psi_1 \circ p(\psi_{2r,2\varepsilon/3}, \psi_{2r,2\varepsilon/3}, \psi_{2r,2\varepsilon/3}) \circ \psi_2$ approximates the product function on $[-r, r]^2$ with accuracy ε. Furthermore, $\mathcal{R}^{\chi_{r,\varepsilon}}_{\mathrm{ReLU}}$ is $\sqrt{32}r$-Lipschitz continuous because $\mathcal{R}^{\psi_{2r,2\varepsilon/3}}_{\mathrm{ReLU}}$ is $4r$-Lipschitz continuous and, hence,

$$\begin{aligned}|\mathcal{R}^{\chi_{r,\varepsilon}}_{\mathrm{ReLU}}(x_1, x_2) - \mathcal{R}^{\chi_{r,\varepsilon}}_{\mathrm{ReLU}}(y_1, y_2)| &\leq 2r\big(|x_1 + x_2 - (y_1 + y_2)| + |x_1 - y_1| + |x_2 - y_2|\big)\\ &\leq \sqrt{32}r\|(x_1 - y_1, x_2 - y_2)\|.\end{aligned}$$

Finally, Proposition 2.1 (items (v) and (vi)) together with Proposition 2.3 (item (iv)) implies that $\mathrm{Cost}_{\mathrm{ReLU},w_0}(\mathrm{pr}, [-r, r]^2, \sqrt{32}r, \varepsilon) \leq 9\mathcal{P}(\psi_{2r,2\varepsilon/3})$. This completes the proof of Proposition 4.2. □

Combining the results from this section, we obtain the following examples of approximable catalogs.

Example 4.3. *Let $q \in (1, \infty)$ and denote by w_q the weight function $(1 + x^q)^{-1}$. For $K \in (0, \infty)$, introduce the K-Lipschitz catalog (augmented by $\mathrm{id}_{\mathbb{R}}$ if $K < 1$)*

$$\mathcal{F}^{\mathrm{Lip}}_K = \big\{f \in C(\mathbb{R}, \mathbb{R})\colon f \text{ is } K\text{-Lipschitz continuous on } \mathbb{R} \text{ and } |f(0)| \leq K\big\} \cup \{\mathrm{id}_{\mathbb{R}}\}$$

and the K-Lipschitz-maximum catalog $\mathcal{F}^{\mathrm{Lip,max}}_K = \mathcal{F}^{\mathrm{Lip}}_K \cup \{\max_d\colon d \in \mathbb{N}\}$. Define prescribed Lipschitz constants and approximation sets

$$L_f = \begin{cases} 1 & \text{if } f = \max_d,\ d \in \mathbb{N},\\ K & \text{if } f \in \mathcal{F}^{\mathrm{Lip}}_K \backslash \{\mathrm{id}_{\mathbb{R}}\},\end{cases} \qquad B_f = \begin{cases} \mathbb{R}^d & \text{if } f = \max_d,\ d \in \mathbb{N},\\ \mathbb{R} & \text{if } f \in \mathcal{F}^{\mathrm{Lip}}_K \backslash \{\mathrm{id}_{\mathbb{R}}\}.\end{cases}$$

Then

(i) $\mathcal{F}^{\mathrm{Lip}}_K$ *is*[1] *a* $[\mathrm{ReLU}, w_q, B, L, \min\{1, 6K\}, (K, 11(6K)^{\frac{q}{q-1}}, 0, \frac{q}{q-1})]$*-approximable catalog*

(ii) *and* $\mathcal{F}^{\mathrm{Lip,max}}_K$ *is a* $[\mathrm{ReLU}, w_q, B, L, \min\{1, 6K\}, (K, 11(6K)^{\frac{q}{q-1}}, 3, \frac{q}{q-1})]$*-approximable catalog.*

Denote by w_0 the weight function that is constantly 1 and introduce the K-Lipschitz-product catalog $\mathcal{F}^{\mathrm{Lip,prod}}_K = \mathcal{F}^{\mathrm{Lip}}_K \cup \{\mathrm{pr}\}$. Let $r, R \in (0, \infty)$ and define new prescribed Lipschitz constants and approximation sets

$$L_f = \begin{cases} 1 & \text{if } f = \max_d,\ d \in \mathbb{N},\\ K & \text{if } f \in \mathcal{F}^{\mathrm{Lip}}_K \backslash \{\mathrm{id}_{\mathbb{R}}\},\\ \sqrt{32}r & \text{if } f = \mathrm{pr},\end{cases} \qquad B_f = \begin{cases} \mathbb{R}^d & \text{if } f = \max_d,\ d \in \mathbb{N},\\ [-R, R] & \text{if } f \in \mathcal{F}^{\mathrm{Lip}}_K \backslash \{\mathrm{id}_{\mathbb{R}}\},\\ [-r, r]^2 & \text{if } f = \mathrm{pr}.\end{cases}$$

Then

(i) $\mathcal{F}^{\mathrm{Lip}}_K$ *is*[2] *a* $[\mathrm{ReLU}, w_0, B, L, \min\{1, KR\}, (K, 7KR, 0, 1)]$*-approximable catalog,*

(ii) $\mathcal{F}^{\mathrm{Lip,max}}_K$ *is a* $[\mathrm{ReLU}, w_0, B, L, \min\{1, KR\}, (K, 7KR, 3, 1)]$*-approximable catalog,*

(iii) *and* $\mathcal{F}^{\mathrm{Lip,prod}}_K$ *is a* $[\mathrm{ReLU}, w_0, B, L, \delta, (K, M, 0, 1)]$*-approximable catalog,*

where $\delta \in (0, 1]$ and $M \in [0, \infty)$ are chosen such that both $3KR\varepsilon^{-1} + 4 \leq M\varepsilon^{-1}$ and $\max\{468, 679 + 720\log_2(r) + 360\log_2(\varepsilon^{-1})\} \leq M\varepsilon^{-1}$ hold for all $\varepsilon \in (0, \delta]$.

[1] Here we use that $\varepsilon \leq \min\{1, 6K\}$ ensures $(6K)^{\frac{q}{q-1}}\varepsilon^{-\frac{q}{q-1}} + 10 \leq 11(6K)^{\frac{q}{q-1}}\varepsilon^{-\frac{q}{q-1}}$. We could also use $\varepsilon = 1$ but then κ_1 becomes larger.

[2] Similarly, we use that $\varepsilon \leq \min\{1, KR\}$ ensures $3KR\varepsilon^{-1} + 4 \leq 7KR\varepsilon^{-1}$.

5 Approximation Results

In this section, we state and prove our main result on the approximability of catalog networks with neural networks. The following lemma is crucial for its proof. It establishes the approximability of the functions $\mathcal{G}_C^{\xi,n}$, $n \in \{1,\dots,\mathcal{D}_C(\xi)\}$, in a catalog network $\xi \in \mathcal{C}_{\mathcal{F}}$ corresponding to an approximable catalog $\mathcal{F}$.

Lemma 5.1. *Assume $a \in C(\mathbb{R},\mathbb{R})$ satisfies the c-identity requirement for some number $c \in [2,\infty)$. Let $\mathcal{F}$ be an $[a,w,B,L,\varepsilon,\kappa]$-approximable catalog with a non-increasing weight function w, and consider a catalog network $\xi \in \mathcal{C}_{\mathcal{F}}^{l_0,\dots,l_{2D}}$ for some $D \in \mathbb{N}$ and $l_0,\dots,l_{2D} \in \mathbb{N}$. Then for all $n \in \{1,\dots,D\}$ and $\delta \in (0,\varepsilon]$, there exists a neural network $\phi \in \mathcal{N}$ with $\mathcal{R}_a^\phi \in C(\mathbb{R}^{l_{2n-1}},\mathbb{R}^{l_{2n}})$ such that*

(i) *$\mathcal{R}_a^\phi$ is $L^{\xi,n}$-Lipschitz continuous on $\mathbb{R}^{l_{2n-1}}$,*

(ii) *$\sup_{x\in\mathbb{B}_B^{\xi,n}} w(\|x\|)\|\mathcal{G}_C^{\xi,n}(x) - \mathcal{R}_a^\phi(x)\| \leq \delta$,*

(iii) *and $\mathcal{P}(\phi) \leq \frac{25}{32}c^4|\kappa_1|^2|l_{2n}|^4|\max\{l_{2n-1},l_{2n}\}|^{2\kappa_2}|\min\{l_{2n-1},l_{2n}\}|^{\kappa_3+2}\delta^{-2\kappa_3}$.*

Proof. Suppose that ξ is given by $[(V_1,b_1,(f_{1,1},\dots,f_{1,k_1})),\dots,(V_D,b_D,(f_{D,1},\dots,f_{D,k_D}))]$ and fix $n \in \{1,\dots,L\}$, $\delta \in (0,\varepsilon]$. The hypothesis that $\mathcal{F}$ is an $[a,w,B,L,\varepsilon,\kappa]$-approximable catalog, where $\kappa = (\kappa_0,\kappa_1,\kappa_2,\kappa_3)$, implies that there exist neural networks $\psi_j \in \mathcal{N}$, $j \in \{1,\dots,k_n\}$, with $\mathcal{R}_a^{\psi_j} \in C(\mathbb{R}^{d_0(f_{n,j})},\mathbb{R}^{d_1(f_{n,j})})$ such that

(I) for all $x,y \in \mathbb{R}^{d_0(f_{n,j})}$ we have $\|\mathcal{R}_a^{\psi_j}(x) - \mathcal{R}_a^{\psi_j}(y)\| \leq L_{f_{n,j}}\|x-y\|$,

(II) for all $x \in B_{f_{n,j}}$ we have $w(\|x\|)\|f_{n,j}(x) - \mathcal{R}_a^{\psi_j}(x)\| \leq \frac{\delta}{\sqrt{k_n}}$,

(III) and $\mathcal{P}(\psi_j) \leq \kappa_1|\max\{d_0(f_{n,j}),d_1(f_{n,j})\}|^{\kappa_2}|k_n|^{\frac{\kappa_3}{2}}\delta^{-\kappa_3}$.

Let $I \in \mathcal{N}$ be such that a fulfills the c-identity requirement with I and let $\phi \in \mathcal{N}$ be given by the parallelization $\phi = p_I(\psi_1,\dots,\psi_{k_n})$. Recall the following piece of notation. For all $j \in \{1,\dots,k_n\}$ and $x \in \mathbb{R}^{l_{2n-1}}$, let $x_{(n,j)} \in \mathbb{R}^{d_0(f_{n,j})}$ be given by

$$x_{(n,j)} = \Big(x_{d_0(f_{n,1})+\cdots+d_0(f_{n,j-1})+1},\dots,x_{d_0(f_{n,1})+\cdots+d_0(f_{n,j})}\Big) \in \mathbb{R}^{d_0(f_{n,j})}.$$

With this notation, we can estimate for all $x,y \in \mathbb{R}^{l_{2n-1}}$,

$$\begin{aligned}\|\mathcal{R}_a^\phi(x) - \mathcal{R}_a^\phi(y)\|^2 &= \sum_{j=1}^{k_n}\|\mathcal{R}_a^{\psi_j}(x_{(n,j)}) - \mathcal{R}_a^{\psi_j}(y_{(n,j)})\|^2 \\ &\leq \sum_{j=1}^{k_n}|L_{f_{n,j}}|^2\|x_{(n,j)} - y_{(n,j)}\|^2 \leq |L^{\xi,n}|^2\|x-y\|^2.\end{aligned} \tag{5.1}$$

Next, we use that w is non-increasing and the definition of $\mathbb{B}_B^{\xi,n}$ to deduce for all $x \in \mathbb{B}_B^{\xi,n}$,

$$\begin{aligned}\|\mathcal{G}_C^{\xi,n}(x) - \mathcal{R}_a^\phi(x)\|^2 &= \sum_{j=1}^{k_n}\|f_{n,j}(x_{(n,j)}) - \mathcal{R}_a^{\psi_j}(x_{(n,j)})\|^2 \\ &\leq \frac{\delta^2}{k_n}\sum_{j=1}^{k_n}|w(\|x_{(n,j)}\|)|^{-2} \leq \delta^2|w(\|x\|)|^{-2}.\end{aligned} \tag{5.2}$$

It remains to estimate the number of parameters $\mathcal{P}(\phi)$. Since $l_{\mathcal{D}(\psi_j)}^{\psi_j} = d_1(f_{n,j}) \leq l_{2n}$ and $\mathcal{D}(\psi_j) \leq \frac{1}{2}\mathcal{P}(\psi_j)$ for all $j \in \{1,\dots,k_n\}$, Proposition 2.5 yields

$$\begin{aligned}\mathcal{P}(\phi) &\leq \frac{1}{2}\Big[\sum_{j=1}^{k_n}c\mathcal{P}(\psi_j) + \frac{cl_{2n}}{2}(cl_{2n}+1)\max_{i\in\{1,\dots,k_n\}}\mathcal{P}(\psi_i)\Big]^2 \\ &\leq \frac{|k_n|^2}{2}\Big[c + \frac{cl_{2n}}{2}(cl_{2n}+1)\Big]^2\Big[\max_{i\in\{1,\dots,k_n\}}\mathcal{P}(\psi_i)\Big]^2 \leq \frac{|k_n|^2}{2}\Big[\tfrac{5}{4}c^2|l_{2n}|^2\Big]^2\Big[\max_{i\in\{1,\dots,k_n\}}\mathcal{P}(\psi_i)\Big]^2,\end{aligned} \tag{5.3}$$

where the last inequality is true because $c \geq 2$. Plugging in item (III) yields

$$\mathcal{P}(\phi) \leq \tfrac{25}{32} c^4 |\kappa_1|^2 |l_{2n}|^4 |\max\{l_{2n-1}, l_{2n}\}|^{2\kappa_2} |k_n|^{\kappa_3+2} \delta^{-2\kappa_3}.$$

Finally, note that we must always have $k_n \leq \min\{l_{2n-1}, l_{2n}\}$, with which we can conclude the proof. □

Remark 5.2. If all functions in a catalog $\mathcal{F}$ have target dimension 1; that is, $d_1(f) = 1$ for all $f \in \mathcal{F}$, then statement (iii) of Lemma 5.1 can be improved to

$$\mathcal{P}(\phi) \leq \tfrac{25}{32} c^4 |\kappa_1|^2 |\max\{l_{2n-1}, l_{2n}\}|^{2\kappa_2} |\min\{l_{2n-1}, l_{2n}\}|^{\kappa_3+2} \delta^{-2\kappa_3}$$

(where we dropped the term $|l_{2n}|^4$). Indeed, in (5.3) we used the estimate $l^{\psi_j}_{\mathcal{D}(\psi_j)} = d_1(f_{n,j}) \leq l_{2n}$ which now becomes $l^{\psi_j}_{\mathcal{D}(\psi_j)} = d_1(f_{n,j}) = 1$, so (5.3) can be improved to

$$\mathcal{P}(\phi) \leq \frac{|k_n|^2}{2} \Big[\tfrac{5}{4} c^2\Big]^2 \Big[\max_{i \in \{1,\dots,k_n\}} \mathcal{P}(\psi_i)\Big]^2.$$

For our main result, we have to introduce a few more concepts. Let $\mathcal{F}$ be a catalog that is approximable on sets $B = (B_f)_{f\in\mathcal{F}}$ with Lipschitz constants $L = (L_f)_{f\in\mathcal{F}} \subseteq [0, \infty)$. Then, for any catalog network

$$\xi = [(V_1, b_1, (f_{1,1}, \dots, f_{1,k_1})), \dots, (V_D, b_D, (f_{D,1}, \dots, f_{D,k_D}))] \in \mathcal{C}_{\mathcal{F}},$$

we define

$$\mathbb{D}^{\xi}_B := \left\{ x \in \mathbb{R}^{\mathcal{I}_C(\xi)} \colon \Big[\text{for all } \{1, \dots, \mathcal{D}_C(\xi)\} \colon \big(\mathcal{A}^{\xi,n}_C \circ \mathcal{G}^{\xi,n-1}_C \circ \cdots \circ \mathcal{A}^{\xi,2}_C \circ \mathcal{G}^{\xi,1}_C \circ \mathcal{A}^{\xi,1}_C\big)(x) \in \mathbb{B}^{\xi,n}_B\Big] \right\}$$

and

$$\mathrm{Lip}_L(\xi) := \prod_{n=1}^{D} L^{\xi,n} \|V_n\|,$$

where $\|\cdot\|$ denotes the operator norm when applied to matrices.

To estimate the approximation error, we need two more quantities. The first one is

$$\mathcal{B}_C(\xi) = \max\big\{1, \|\mathcal{A}^{\xi,1}_C(0)\|, \dots, \|\mathcal{A}^{\xi,\mathcal{D}_C(\xi)}_C(0)\|\big\},$$

which simply measures the maximal norm of the inhomogeneous parts of the affine transformations (capped from below by 1). When using weight functions of the type $w_q(x) = (1 + x^q)^{-1}$, functions in the catalog are approximated better close to the origin. The quantity $\mathcal{B}_C(\xi)$ is used to control how far away one is from the region where one has the best approximation. However, this becomes redundant if the weight function is constant as Corollary 5.8 will show. The last quantity we need is

$$\mathcal{T}_L(\xi) = \max_{\substack{m,n\in\mathbb{N}_0,\\ m<n\leq D+1}} \Big[\max\big\{1, \|V_{\min\{n,D\}}\|\big\} \max\Big\{1, L^{\xi,\max\{m,1\}}\Big\} \Big(\prod_{j=m+1}^{n-1} L^{\xi,j} \|V_j\|\Big)\Big].$$

Essentially, this is the maximum of all Lipschitz constants of layers $m+1$ to n. Its precise use will become clear in the proof of the following result.

Theorem 5.3. *Suppose $a \in C(\mathbb{R}, \mathbb{R})$ fulfills the c-identity requirement for some number $c \geq 2$ and let w be a non-increasing weight function with order of growth at most (s_1, s_2) for some $s_1 \in [1, \infty)$ and $s_2 \in [0, \infty)$. Consider a catalog network $\xi \in \mathcal{C}_{\mathcal{F}}$ for an $[a, w, B, L, \varepsilon, \kappa]$-approximable catalog $\mathcal{F}$. Then there exists a neural network architecture $\phi \in \mathcal{N}$ with a-realization $\mathcal{R}^{\phi}_a \in C(\mathbb{R}^{\mathcal{I}_C(\xi)}, \mathbb{R}^{\mathcal{O}_C(\xi)})$ such that*

(i) *$\mathcal{R}^{\phi}_a$ is $\mathrm{Lip}_L(\xi)$-Lipschitz continuous on $\mathbb{R}^{\mathcal{I}_C(\xi)}$,*

(ii) *$\sup_{x\in\mathbb{D}^{\xi}_B} w(\|x\|)\|\mathcal{R}^{\phi}_a(x) - \mathcal{R}^{\xi}_C(x)\| \leq \varepsilon$, and*

(iii) $\mathcal{P}(\phi) \le C|\mathcal{B}_C(\xi)|^{2(r-\kappa_3)}|\mathcal{T}_L(\xi)|^{2r}|\mathcal{D}_C(\xi)|^{2r+1}|\mathcal{W}_C(\xi)|^{2\kappa_2+r+7}\varepsilon^{-2\kappa_3}$
for $r = \kappa_3(s_2+1)$ *and* $C = 43 \cdot 2^{4(r-\kappa_3)-5}c^5|\kappa_0|^{2(r-\kappa_3)}|\kappa_1|^2|s_1|^{2\kappa_3}$.

Remark 5.4. The conclusion of Theorem 5.3 could be written more concisely as

$$\mathrm{Cost}_{a,w}(\mathcal{R}_C^\xi, \mathbb{D}_B^\xi, \mathrm{Lip}_L(\xi), \varepsilon) \le C|\mathcal{B}_C(\xi)|^{2(r-\kappa_3)}|\mathcal{T}_L(\xi)|^{2r}|\mathcal{D}_C(\xi)|^{2r+1}|\mathcal{W}_C(\xi)|^{2\kappa_2+r+7}\varepsilon^{-2\kappa_3}.$$

In particular, for any given $M \in [1,\infty)$ and $k \in [0,\infty)$, the set of functions

$$\left\{\mathcal{R}_C^\xi\colon \left[\begin{array}{c} \xi \in \mathcal{C}_\mathcal{F} \text{ with } \|\mathcal{R}_C^\xi(0)\| \le M \text{ and} \\ \max\{\mathcal{B}_C(\xi), \mathcal{T}_L(\xi), \mathcal{D}_C(\xi), \mathcal{W}_C(\xi)\} \le |\max\{\mathcal{I}_C(\xi), \mathcal{O}_C(\xi)\}|^k \end{array}\right]\right\}$$

is an $[a, w, \mathbb{D}_B^{(\cdot)}, \mathrm{Lip}_L(\cdot), \varepsilon, (M, C, k(2\kappa_2 - 2\kappa_3 + 7r + 8), 2\kappa_3)]$-approximable catalog.

Proof of Theorem 5.3. We split the proof into two parts. In the first half, we construct a candidate neural network for our approximation and bound the approximation error. In the second half, we analyze the number of parameters. Assume that $\xi \in \mathcal{C}_\mathcal{F}^{l_0,\dots,l_{2D}}$ is given by $\xi = [(V_1, b_1, (f_{1,1}, \dots, f_{1,k_1})), \dots, (V_D, b_D, (f_{D,1}, \dots, f_{D,k_D}))]$. Introduce the short-hand $G_n \in C(\mathbb{R}^{l_0}, \mathbb{R}^{l_{2n}})$, $n \in \{0, \dots, D\}$, given by

$$G_n = \mathcal{G}_C^{\xi,n} \circ \mathcal{A}_C^{\xi,n} \circ \mathcal{G}_C^{\xi,n-1} \circ \cdots \circ \mathcal{G}_C^{\xi,1} \circ \mathcal{A}_C^{\xi,1}$$

for $n \in \{1, \dots, D\}$ and $G_0 = \mathrm{id}_{\mathbb{R}^{l_0}}$. Before we consider any specific neural networks, let us work on a general bound for $\mathcal{A}_C^{\xi,n} \circ G_{n-1}$. We prove by induction over n that for all $n \in \{1, \dots, D\}$ and $x \in \mathbb{D}_B^\xi$

$$\begin{aligned} \|(\mathcal{A}_C^{\xi,n} \circ G_{n-1})(x)\| &\le \|V_n\|\Big(\prod_{j=1}^{n-1} L^{\xi,j}\|V_j\|\Big)\|x\| + \|b_n\| \\ &\quad + \sum_{m=1}^{n-1} \|V_n\|\Big(\prod_{j=m+1}^{n-1} L^{\xi,j}\|V_j\|\Big)\big(L^{\xi,m}\|b_m\| + \|\mathcal{G}_C^{\xi,m}(0)\|\big). \end{aligned} \tag{5.4}$$

The base case $n = 1$ reduces to $\|\mathcal{A}_C^{\xi,1}(x)\| \le \|V_1\|\|x\| + \|b_1\|$. For the induction step, suppose the claim is true for a given $n \in \{1, \dots, D-1\}$. Then Lemma 3.5 implies for all $x \in \mathbb{D}_B^\xi$

$$\begin{aligned} \|(\mathcal{A}_C^{\xi,n+1} \circ G_n)(x)\| &\le \|V_{n+1}\|\|(\mathcal{G}_C^{\xi,n} \circ \mathcal{A}_C^{\xi,n} \circ G_{n-1})(x)\| + \|b_{n+1}\| \\ &\le \|V_{n+1}\|(L^{\xi,n}\|(\mathcal{A}_C^{\xi,n} \circ G_{n-1})(x)\| + \|\mathcal{G}_C^{\xi,n}(0)\|) + \|b_{n+1}\|. \end{aligned}$$

For this step, it was crucial in the definition of an approximable catalog to require the sets B_f to contain the origin. Plugging the induction hypothesis into $\|(\mathcal{A}_C^{\xi,n} \circ G_{n-1})(x)\|$, we readily obtain the formula in (5.4). Next, observe that for all $n \in \{1, \dots, D\}$

$$\|\mathcal{G}_C^{\xi,n}(0)\|^2 = \sum_{j=1}^{k_n} \|f_{n,j}(0)\|^2 \le k_n|\kappa_0|^2 \le |\mathcal{W}_C(\xi)||\kappa_0|^2. \tag{5.5}$$

Hence, (5.4) yields for all $n \in \{1, \dots, D\}$ and $x \in \mathbb{D}_B^\xi$

$$\begin{aligned} \|(\mathcal{A}_C^{\xi,n} \circ G_{n-1})(x)\| &\le |\mathcal{T}_L(\xi)|\|x\| + \mathcal{B}_C(\xi) + D|\mathcal{T}_L(\xi)|\big(\mathcal{B}_C(\xi) + |\kappa_0|\sqrt{\mathcal{W}_C(\xi)}\big) \\ &\le 4|\kappa_0||\mathcal{B}_C(\xi)||\mathcal{D}_C(\xi)|\sqrt{\mathcal{W}_C(\xi)}|\mathcal{T}_L(\xi)|\max\{1, \|x\|\}. \end{aligned} \tag{5.6}$$

Now we start constructing our candidate neural network. Lemma 5.1 shows that for all $\delta \in (0, \varepsilon]$, $n \in \{1, \dots, D\}$ there exists $\psi_{\delta,n} \in \mathcal{N}$ with $\mathcal{R}_a^{\psi_{\delta,n}} \in C(\mathbb{R}^{l_{2n-1}}, \mathbb{R}^{l_{2n}})$ such that

(I) for all $x, y \in \mathbb{R}^{l_{2n-1}}$ we have $\|\mathcal{R}_a^{\psi_{\delta,n}}(x) - \mathcal{R}_a^{\psi_{\delta,n}}(y)\| \le L^{\xi,n}\|x - y\|$,

(II) for all $x \in \mathbb{B}_B^{\xi,n}$ we have $w(\|x\|)\|\mathcal{G}_C^{\xi,n}(x) - \mathcal{R}_a^{\psi_{\delta,n}}(x)\| \le \delta$,

(III) and $\mathcal{P}(\psi_{\delta,n}) \leq \frac{25}{32} c^4 |\kappa_1|^2 |\mathcal{W}_C(\xi)|^{2\kappa_2+\kappa_3+6} \delta^{-2\kappa_3}$.

Moreover, since each $\mathcal{A}_C^{\xi,n}$ is an affine function, there exist unique $\chi_n \in \mathcal{N}$, $n \in \{1,\dots,D\}$, of depth 1 with $\mathcal{R}_a^{\chi_n} \in C(\mathbb{R}^{l_{2n-2}}, \mathbb{R}^{l_{2n-1}})$ given by $\mathcal{R}_a^{\chi_n} = \mathcal{A}_C^{\xi,n}$. Let $\varphi_{\delta,n} \in \mathcal{N}$ be given by $\varphi_{\delta,n} = \psi_{\delta,n} \circ \chi_n \circ \cdots \circ \psi_{\delta,1} \circ \chi_1$. With $n = D$, this will be our candidate neural network for sufficiently small δ. Let us verify that it does the job in terms of the approximation precision. To do so, we prove by induction over n that for all $n \in \{1,\dots,D\}$, $\delta \in (0,\varepsilon]$, $x \in \mathbb{D}_B^\xi$

$$\|\mathcal{R}_a^{\varphi_{\delta,n}}(x) - G_n(x)\| \leq \delta \sum_{m=1}^{n} \frac{\prod_{j=m+1}^{n} L^{\xi,j} \|V_j\|}{w(\|\mathcal{A}_C^{\xi,m} \circ G_{m-1}(x)\|)}. \tag{5.7}$$

The base case $n = 1$ holds by the approximation property of $\psi_{\delta,1}$ and the fact that $\mathcal{A}_C^{\xi,1}(x) \in \mathbb{B}_B^{\xi,1}$ whenever $x \in \mathbb{D}_B^\xi$. For the induction step, suppose the claim is true for a given $n \in \{1,\dots,D-1\}$. By the Lipschitz and the approximation properties of $\psi_{\delta,n+1}$, we obtain for all $\delta \in (0,\varepsilon]$, $x \in \mathbb{D}_B^\xi$

$$\begin{aligned}\|\mathcal{R}_a^{\varphi_{\delta,n+1}}(x) - G_{n+1}(x)\| &= \|\mathcal{R}_a^{\psi_{\delta,n+1}}((\mathcal{A}_C^{\xi,n+1} \circ \mathcal{R}_a^{\varphi_{\delta,n}}(x)) - \mathcal{G}_C^{\xi,n+1}((\mathcal{A}_C^{\xi,n+1} \circ G_n)(x))\| \\ &\leq |L^{\xi,n+1}| \|V_{n+1}\| \|\mathcal{R}_a^{\varphi_{\delta,n}}(x) - G_n(x)\| + \frac{\delta}{w(\|(\mathcal{A}_C^{\xi,n+1} \circ G_n)(x)\|)},\end{aligned}$$

where we used that $(\mathcal{A}_C^{\xi,n+1} \circ G_n)(x) \in \mathbb{B}_B^{\xi,n+1}$ whenever $x \in \mathbb{D}_B^\xi$. Plugging the induction hypothesis into $\|\mathcal{R}_a^{\varphi_{\delta,n}}(x) - G_n(x)\|$, we readily obtain the formula in (5.7). Now we combine (5.6), (5.7), and the hypothesis that w has order of growth at most (s_1, s_2) to find for all $\delta \in (0,\varepsilon]$, $x \in \mathbb{D}_B^\xi$

$$\begin{aligned}\|\mathcal{R}_a^{\varphi_{\delta,D}}(x) - G_D(x)\| &\leq \frac{\delta D |\mathcal{T}_L(\xi)|}{w\big(4|\kappa_0||\mathcal{B}_C(\xi)||\mathcal{D}_C(\xi)|\sqrt{\mathcal{W}_C(\xi)}|\mathcal{T}_L(\xi)| \max\{1, \|x\|\}\big)} \\ &\leq s_1 \delta 4^{s_2} |\kappa_0|^{s_2} |\mathcal{B}_C(\xi)|^{s_2} |\mathcal{D}_C(\xi)|^{s_2+1} |\mathcal{W}_C(\xi)|^{\frac{s_2}{2}} |\mathcal{T}_L(\xi)|^{s_2+1} |w(\|x\|)|^{-1}.\end{aligned} \tag{5.8}$$

This essentially finishes the approximation part of the proof if we pick δ appropriately. That $\mathcal{R}_a^{\varphi_{\delta,D}}$ is $\mathrm{Lip}_L(\xi)$-Lipschitz continuous follows from the fact that the concatenation of Lipschitz functions is again Lipschitz with constant equal to the product of the original Lipschitz constants. It remains to estimate the number of parameters in the constructed neural network. Actually, let us make a slight modification. Pick $I \in \mathcal{N}$ for which a fulfills the c-identity requirement and let $I_d \in \mathcal{N}$, $d \in \mathbb{N}$, denote the d-fold parallelization of I. Then, let $\rho_{\delta,n} \in \mathcal{N}$, $n \in \{1,\dots,D\}$, $\delta \in (0,\varepsilon]$, be given by $\rho_{\delta,n} = I_{l_{2n}} \circ \psi_{\delta,n} \circ I_{l_{2n-1}}$. Solely for approximation purposes, one could simply work with $\psi_{\delta,n}$ instead of $\rho_{\delta,n}$, but for the latter we have better control on the number of parameters once we start concatenating the layers. Combining Corollary 2.6 and the bound for $\mathcal{P}(\psi_{\delta,n})$, we see that for all $\delta \in (0,\varepsilon]$ and $n \in \{1,\dots,D\}$

$$\begin{aligned}\mathcal{P}(\rho_{\delta,n}) &\leq \tfrac{25}{32} c^5 |\kappa_1|^2 |\mathcal{W}_C(\xi)|^{2\kappa_2+\kappa_3+7} \delta^{-2\kappa_3} + 3c^2 |\mathcal{W}_C(\xi)|^2 \\ &\leq \tfrac{37}{32} c^5 |\kappa_1|^2 |\mathcal{W}_C(\xi)|^{2\kappa_2+\kappa_3+7} \delta^{-2\kappa_3},\end{aligned} \tag{5.9}$$

where we used $c \geq 2$. Next, note that Proposition 2.3 implies $l_1^{I_d} \leq cd$ for all $d \in \mathbb{N}$. Thus, Proposition 2.1 yields for all $\delta \in (0,\varepsilon]$ and $n \in \{1,\dots,D\}$

$$\begin{aligned}\mathcal{P}(\rho_{\delta,n} \circ \chi_n) &\leq \mathcal{P}(\rho_{\delta,n}) + \mathcal{P}(\chi_n) + c l_{2n-1} l_{2n-2} - l_{2n-1}(l_{2n-2}+1) \\ &\leq \mathcal{P}(\rho_{\delta,n}) + c|\mathcal{W}_C(\xi)|^2.\end{aligned} \tag{5.10}$$

Furthermore, we prove by induction over n that for all $n \in \{1,\dots,D\}$, $\delta \in (0,\varepsilon]$

$$\mathcal{P}(\rho_{\delta,n} \circ \chi_n \circ \cdots \circ \rho_{\delta,1} \circ \chi_1) \leq (n-1)c^2 |\mathcal{W}_C(\xi)|^2 + \sum_{j=1}^{n} \mathcal{P}(\rho_{\delta,j} \circ \chi_j). \tag{5.11}$$

The base case $n = 1$ is trivially satisfied. For the induction step, suppose the claim is true for a given $n \in \{1, \dots, D-1\}$. Then Proposition 2.1 and the induction hypothesis show for all $\delta \in (0, \varepsilon]$

$$\begin{aligned}\mathcal{P}(\rho_{\delta,n+1} \circ \chi_{n+1} \circ \cdots \circ \rho_{\delta,1} \circ \chi_1) &\leq \mathcal{P}(\rho_{\delta,n+1} \circ \chi_{n+1}) + \mathcal{P}(\rho_{\delta,n} \circ \chi_n \circ \cdots \circ \rho_{\delta,1} \circ \chi_1) + c^2 l_{2(n+1)-1} l_{2n} \\ &\leq nc^2 |\mathcal{W}_C(\xi)|^2 + \sum_{j=1}^{n+1} \mathcal{P}(\rho_{\delta,j} \circ \chi_j),\end{aligned}$$

which finishes the induction. Now we combine (5.10) and (5.11) to obtain for all $\delta \in (0, \varepsilon]$

$$\mathcal{P}(\rho_{\delta,D} \circ \chi_D \circ \cdots \circ \rho_{\delta,1} \circ \chi_1) \leq \tfrac{3}{2} c^2 |\mathcal{D}_C(\xi)||\mathcal{W}_C(\xi)|^2 + \sum_{j=1}^{D} \mathcal{P}(\rho_{\delta,j}),$$

where we used $c \geq 2$ again. Plugging (5.9) into this inequality yields for all $\delta \in (0, \varepsilon]$

$$\begin{aligned}\mathcal{P}(\rho_{\delta,D} \circ \chi_D \circ \cdots \circ \rho_{\delta,1} \circ \chi_1) &\leq \tfrac{3}{2} c^2 |\mathcal{D}_C(\xi)||\mathcal{W}_C(\xi)|^2 + D\big(\tfrac{37}{32} c^5 |\kappa_1|^2 |\mathcal{W}_C(\xi)|^{2\kappa_2+\kappa_3+7} \delta^{-2\kappa_3}\big) \\ &\leq \tfrac{43}{32} c^5 |\kappa_1|^2 |\mathcal{D}_C(\xi)||\mathcal{W}_C(\xi)|^{2\kappa_2+\kappa_3+7} \delta^{-2\kappa_3}.\end{aligned} \tag{5.12}$$

This finishes the parameter estimation. We conclude by gathering everything we have proved so far. Motivated by (5.6), let $\eta \in (0, \varepsilon]$ be given by

$$\eta = \Big[4^{s_2} |s_1| |\kappa_0|^{s_2} |\mathcal{B}_C(\xi)|^{s_2} |\mathcal{D}_C(\xi)|^{s_2+1} |\mathcal{W}_C(\xi)|^{\frac{s_2}{2}} |\mathcal{T}_L(\xi)|^{s_2+1}\Big]^{-1} \varepsilon$$

and let $\phi \in \mathcal{N}$ be given by $\phi = \rho_{\eta,D} \circ \chi_D \circ \cdots \circ \rho_{\eta,1} \circ \chi_1$. Note that $\mathcal{R}_a^\phi = \mathcal{R}_a^{\varphi_{\eta,D}}$ and $G_D = \mathcal{R}_C^\xi$. Thus, (5.8) translates to

$$\sup_{x \in \mathbb{D}_B^\xi} w(\|x\|) \|\mathcal{R}_a^\phi(x) - \mathcal{R}_C^\xi(x)\| \leq \varepsilon.$$

Moreover, (5.12) implies

$$\begin{aligned}\mathcal{P}(\phi) \leq 43 \cdot 2^{4\kappa_3 s_2 - 5} c^5 |\kappa_0|^{2\kappa_3 s_2} |\kappa_1|^2 |s_1|^{2\kappa_3} |\mathcal{B}_C(\xi)|^{2\kappa_3 s_2} |\mathcal{T}_L(\xi)|^{2\kappa_3(s_2+1)} \\ \cdot |\mathcal{D}_C(\xi)|^{2\kappa_3(s_2+1)+1} |\mathcal{W}_C(\xi)|^{2\kappa_2+\kappa_3(s_2+1)+7} \varepsilon^{-2\kappa_3},\end{aligned}$$

which completes the proof of Theorem 5.3. □

Remark 5.5. By Remark 5.2, if $d_1(f) = 1$ for all $f \in \mathcal{F}$, then item (III) in the proof of Theorem 5.3 can be improved to

$$\mathcal{P}(\psi_{\delta,n}) \leq \tfrac{25}{32} c^4 |\kappa_1|^2 |\mathcal{W}_C(\xi)|^{4\kappa_2+\kappa_3+2} \delta^{-2\kappa_3}$$

(the exponent of $\mathcal{W}_C(\xi)$ changed). As a consequence, item (iii) of Theorem 5.3 can be improved to

$$\mathcal{P}(\phi) \leq C |\mathcal{B}_C(\xi)|^{2(r-\kappa_3)} |\mathcal{T}_L(\xi)|^{2r} |\mathcal{D}_C(\xi)|^{2r+1} |\mathcal{W}_C(\xi)|^{2\kappa_2+r+3} \varepsilon^{-2\kappa_3}.$$

Remark 5.6. A careful inspection of the proof of Theorem 5.3 reveals that it does not only work with the Euclidean norm. For instance, one could also also work with the ∞-norm. One only needs to replace the property $\|x\|^2 = \sum_{j=1}^n \|x_{(j)}\|^2$ of the Euclidean norm in (3.1), (5.1), (5.2), and (5.5) with the corresponding property $\|x\|^2 = \max_{j \in \{1,\dots,n\}} \|x_{(j)}\|^2$ of the ∞-norm and note that (5.2) is still true because $\|x_{(j)}\| \leq \|x\|$ holds for the ∞-norm, as well.

One activation function we know fulfills the identity requirement is the ReLU function. Our main result recast for ReLU activation and the weight function $w_q(x) = (1 + x^q)^{-1}$ reads as follows:

Corollary 5.7. *Consider the weight function $w_q(x) = (1 + x^q)^{-1}$ for some $q \in (0, \infty)$. Let $\xi \in \mathcal{C}_\mathcal{F}$ be a catalog network for an $[\mathrm{ReLU}, w_q, B, L, \varepsilon, \kappa]$-approximable catalog. Then there exists a neural network architecture $\phi \in \mathcal{N}$ with ReLU-realization $\mathcal{R}_{\mathrm{ReLU}}^\phi \in C(\mathbb{R}^{\mathcal{I}_C(\xi)}, \mathbb{R}^{\mathcal{O}_C(\xi)})$ such that*

(i) $\sup_{x \in \mathbb{D}_B^\xi} (1 + \|x\|^q)^{-1} \|\mathcal{R}_{\mathrm{ReLU}}^\phi(x) - \mathcal{R}_C^\xi(x)\| \leq \varepsilon$, *and*

(ii) $\mathcal{P}(\phi) \leq C|\mathcal{B}_C(\xi)|^{2(r-\kappa_3)}|\mathcal{T}_L(\xi)|^{2r}|\mathcal{D}_C(\xi)|^{2r+1}|\mathcal{W}_C(\xi)|^{2\kappa_2+r+7}\varepsilon^{-2\kappa_3}$,
for $r = \kappa_3(q+1)$ *and* $C = 43 \cdot 4^{2r-\kappa_3}|\kappa_0|^{2(r-\kappa_3)}|\kappa_1|^2$.

Noteworthy is also that for the weight function $w_0 \equiv 1$, the parameter estimate in Theorem 5.3 simplifies considerably:

Corollary 5.8. *Let* $\xi \in \mathcal{C}_{\mathcal{F}}$ *be a catalog network for an* $[\mathrm{ReLU}, w_0, B, L, \varepsilon, \kappa]$*-approximable catalog. Then there exists a neural network architecture* $\phi \in \mathcal{N}$ *with ReLU-realization* $\mathcal{R}^{\phi}_{\mathrm{ReLU}} \in C(\mathbb{R}^{\mathcal{I}_C(\xi)}, \mathbb{R}^{\mathcal{O}_C(\xi)})$ *such that*

(i) $\sup_{x \in \mathbb{D}^{\xi}_B} \|\mathcal{R}^{\phi}_{\mathrm{ReLU}}(x) - \mathcal{R}^{\xi}_C(x)\| \leq \varepsilon$, *and*

(ii) $\mathcal{P}(\phi) \leq 43|\kappa_1|^2|\mathcal{T}_L(\xi)|^{2\kappa_3}|\mathcal{D}_C(\xi)|^{2\kappa_3+1}|\mathcal{W}_C(\xi)|^{2\kappa_2+\kappa_3+7}\varepsilon^{-2\kappa_3}$.

We conclude the section by noting that we could have specified the characteristic inequality of Definition 3.4 as

$$\mathrm{Cost}_{a,w}(f, B_f, L_f, \delta) \leq \kappa_1 |\max\{d_0(f), d_1(f)\}|^{\kappa_2} |\log_2(\delta^{-1})|^{\kappa_3} \tag{5.13}$$

by using a log-term $|\log_2(\delta^{-1})|^{\kappa_3}$ instead of $\delta^{-\kappa_3}$. With this alternative definition, almost all arguments of Section 5 go through and one obtains a version of Lemma 5.1 with statement (iii) modified to

$$\mathcal{P}(\phi) \leq \tfrac{25}{32}c^4|\kappa_1|^2|l_{2n}|^4 |\max\{l_{2n-1}, l_{2n}\}|^{2\kappa_2} |\min\{l_{2n-1}, l_{2n}\}|^2 \big|\log_2\big(\tfrac{|\min\{l_{2n-1}, l_{2n}\}|^{1/2}}{\delta}\big)\big|^{2\kappa_3}.$$

If, subsequently, one replaces (5.9) with

$$\mathcal{P}(\rho_{\delta,n}) \leq \tfrac{37}{32}c^5|\kappa_1|^2|\mathcal{W}_C(\xi)|^{2\kappa_2+7}\big|\log_2\big(|\mathcal{W}_C(\xi)|^{1/2}\delta^{-1}\big)\big|^{2\kappa_3}$$

and (5.12) with

$$\mathcal{P}(\rho_{\delta,D} \circ \chi_D \circ \cdots \circ \rho_{\delta,1} \circ \chi_1) \leq \tfrac{43}{32}c^5|\kappa_1|^2|\mathcal{D}_C(\xi)||\mathcal{W}_C(\xi)|^{2\kappa_2+7}\big|\log_2\big(|\mathcal{W}_C(\xi)|^{1/2}\delta^{-1}\big)\big|^{2\kappa_3},$$

one obtains the following version of Theorem 5.3:

Theorem 5.9. *Assume* $a \in C(\mathbb{R}, \mathbb{R})$ *satisfies the* c*-identity requirement for some number* $c \geq 2$*, and let* w *be a non-increasing weight function with order of growth at most* (s_1, s_2) *for some* $s_1 \in [1, \infty)$ *and* $s_2 \in [0, \infty)$*. Consider a catalog network* $\xi \in \mathcal{C}_{\mathcal{F}}$ *for an* $[a, w, B, L, \varepsilon, \kappa]$*-approximable catalog* $\mathcal{F}$*, where the approximation cost is given by* (5.13) *and* $\varepsilon \leq 1/2$*. Then there exists a neural network architecture* $\phi \in \mathcal{N}$ *with* a*-realization* $\mathcal{R}^{\phi}_a \in C(\mathbb{R}^{\mathcal{I}_C(\xi)}, \mathbb{R}^{\mathcal{O}_C(\xi)})$ *such that*

(i) $\mathcal{R}^{\phi}_a$ *is* $\mathrm{Lip}_L(\xi)$*-Lipschitz continuous on* $\mathbb{R}^{\mathcal{I}_C(\xi)}$,

(ii) $\sup_{x \in \mathbb{D}^{\xi}_B} w(\|x\|)\|\mathcal{R}^{\phi}_a(x) - \mathcal{R}^{\xi}_C(x)\| \leq \varepsilon$, *and*

(iii) $\mathcal{P}(\phi) \leq C_1|\mathcal{D}_C(\xi)||\mathcal{W}_C(\xi)|^{2\kappa_2+7}\big|\log_2\big(C_2|\mathcal{B}_C(\xi)|^{s_2}|\mathcal{D}_C(\xi)|^{s_2+1}|\mathcal{W}_C(\xi)|^{\frac{s_2+1}{2}}|\mathcal{T}_L(\xi)|^{s_2+1}\varepsilon^{-1}\big)\big|^{2\kappa_3}$
for $C_1 = \frac{43}{32}c^5|\kappa_1|^2$ *and* $C_2 = 4^{s_2}|s_1||\kappa_0|^{s_2}$.

The next result is the analogue of Corollary 5.8 for the log-modification (5.13) of the approximation cost.

Corollary 5.10. *Consider the weight function* $w_0 \equiv 1$*, and let* $\xi \in \mathcal{C}_{\mathcal{F}}$ *be a catalog network for an* $[\mathrm{ReLU}, w_0, B, L, \varepsilon, \kappa]$*-approximable catalog* $\mathcal{F}$*, where the approximation cost is given by* (5.13) *and* $\varepsilon \leq 1/2$*. Then there exists a neural network architecture* $\phi \in \mathcal{N}$ *with ReLU-realization* $\mathcal{R}^{\phi}_{\mathrm{ReLU}} \in C(\mathbb{R}^{\mathcal{I}_C(\xi)}, \mathbb{R}^{\mathcal{O}_C(\xi)})$ *such that*

(i) $\sup_{x \in \mathbb{D}^{\xi}_B} \|\mathcal{R}^{\phi}_{\mathrm{ReLU}}(x) - \mathcal{R}^{\xi}_C(x)\| \leq \varepsilon$, *and*

(ii) $\mathcal{P}(\phi) \leq 43|\kappa_1|^2|\mathcal{D}_C(\xi)||\mathcal{W}_C(\xi)|^{2\kappa_2+7}\big|\log_2\big(|\mathcal{D}_C(\xi)||\mathcal{W}_C(\xi)|^{\frac{1}{2}}|\mathcal{T}_L(\xi)|\varepsilon^{-1}\big)\big|^{2\kappa_3}$.

The log-modification (5.13) of the approximation cost is useful when considering catalogs consisting of functions that can be approximated with a number of parameters growing like $\log_2(\varepsilon^{-1})$ in the accuracy ε. With the original definition of the approximation cost, one would obtain an estimate of the form $M\varepsilon^{-\kappa_3}$ for $\log_2(\varepsilon^{-1})$, which is rather rough for small ε; see Propositions 6.4 and 6.5 below. The following example is a consequence of Proposition 4.2.

Example 5.11. Consider the weight function $w_0 \equiv 1$ and let $\mathcal{F}^{\mathrm{prod}} = \{\mathrm{id}_{\mathbb{R}}, \mathrm{pr}\}$ be the product catalog. Fix $r \in (0,\infty)$, and consider the Lipschitz constants and approximation sets

$$L_f = \begin{cases} 1 & \text{if } f = \mathrm{id}_{\mathbb{R}}, \\ \sqrt{32}r & \text{if } f = \mathrm{pr}, \end{cases} \qquad B_f = \begin{cases} \mathbb{R} & \text{if } f = \mathrm{id}_{\mathbb{R}}, \\ [-r,r]^2 & \text{if } f = \mathrm{pr}. \end{cases}$$

Then $\mathcal{F}^{\mathrm{prod}}$ is a $[\mathrm{ReLU}, w_0, B, L, \delta, (1, M, 0, 1)]$-approximable catalog if the approximation cost is measured by (5.13), and $\delta \in (0,1]$ and $M \in [0,\infty)$ are such that $\max\{468, 679+720\log_2(r)+360\log_2(\varepsilon^{-1})\} \leq M\log_2(\varepsilon^{-1})$ for all $\varepsilon \in (0,\delta]$.

6 Overcoming the curse of dimensionality

In this section, we apply the theory of catalog networks to show that different high-dimensional functions admit a ReLU neural network approximation without the curse of dimensionality. We concentrate on catalogs containing one-dimensional Lipschitz functions, maximum functions, and the product function in two dimensions. Building on these, we construct families of functions indexed by the dimension of their domain that are of the same form for each dimension. For instance, in the first example below, we consider the sum of d Lipschitz functions for $d \in \mathbb{N}$.

Proposition 6.1. *Fix $K, R \in (0,\infty)$, and let $g_d\colon \mathbb{R} \to \mathbb{R}$, $d \in \mathbb{N}$, be K-Lipschitz continuous on $\mathbb{R}$ with $|g_d(0)| \leq K$. Define $f_d\colon \mathbb{R}^d \to \mathbb{R}$ by $f_d(x) = \sum_{k=1}^d g_k(x_k)$. Then for all $d \in \mathbb{N}$ and $\varepsilon \in (0, \min\{1, KR\}]$, there exists $\phi \in \mathcal{N}$ with $\mathcal{R}^\phi_{\mathrm{ReLU}} \in C(\mathbb{R}^d, \mathbb{R})$ such that*

(i) $\sup_{x\in[-R,R]^d} |f_d(x) - \mathcal{R}^\phi_{\mathrm{ReLU}}(x)| \leq \varepsilon$,

(ii) *and* $\mathcal{P}(\phi) \leq \frac{2}{11}10^5 \max\{1, K^4\}R^2 d^6 \varepsilon^{-2}$.

Proof. Let $\mathcal{F} = \mathcal{F}^{\mathrm{Lip}}_K$ be the K-Lipschitz catalog and suppose $L = (L_f)_{f\in\mathcal{F}}$ and $B = (B_f)_{f\in\mathcal{F}}$ are defined as for item (i) in Example 4.3. Let $V_d \in \mathbb{R}^{1\times d}$, $d \in \mathbb{N}$, be the matrix $V_d = (1 \ \cdots \ 1)$ with all entries 1 and let $\xi_d \in \mathcal{C}_{\mathcal{F}}$, $d \in \mathbb{N}$, be given by $\xi_d = [(\mathrm{id}_{\mathbb{R}^d}, 0, (g_1, \ldots, g_d)), (V_d, 0, \mathrm{id}_{\mathbb{R}})]$. Then $\mathcal{R}^{\xi_d}_C = f_d$, $\mathcal{D}_C(\xi_d) = 2$, $\mathcal{W}_C(\xi_d) = d$, and $\mathcal{T}_L(\xi_d) \leq d\max\{1, K^2\}$. Moreover, $\mathbb{D}^{\xi_d}_B = \mathbb{B}^{\xi_d,1}_B = [-R,R]^d$ because $\mathbb{B}^{\xi_d,2}_B = B_{\mathrm{id}_{\mathbb{R}}} = \mathbb{R}$. Thus, Example 4.3, Remark 5.5, and Corollary 5.8 yield Proposition 6.1. □

While the factor 10^5 may seem large, the growth of $10^5 d^6$ in the dimension is much slower than exponential growth, say 2^d, as soon as the dimension exceeds roughly 50. Thus, for large dimension, our estimates are better than those of other approximation schemes suffering from the curse of dimensionality.

Note that the approximating networks constructed in the proof of Proposition 6.1 have fixed depth independent of both the dimension and the accuracy since we approximate all Lipschitz functions in the catalog with networks with a single hidden layer. This illustrates that it may happen that the networks provided by Theorem 5.3 only grow in width but not in depth as the dimension increases.

Proposition 6.2. *Fix $K, R \in (0,\infty)$, and let $g_d\colon \mathbb{R} \to \mathbb{R}$, $d \in \mathbb{N}$, be K-Lipschitz continuous with $|g_d(0)| \leq K$. Consider the functions $f_d\colon \mathbb{R}^d \to \mathbb{R}$, given by $f_d(x) = \max\{g_1(x_1), \ldots, g_d(x_d)\}$. Then for all $d \in \mathbb{N}$ and $\varepsilon \in (0, \min\{1, KR\}]$, there exists $\phi \in \mathcal{N}$ with $\mathcal{R}^\phi_{\mathrm{ReLU}} \in C(\mathbb{R}^d, \mathbb{R})$ such that*

(i) $\sup_{x\in[-R,R]^d} |f_d(x) - \mathcal{R}^\phi_{\mathrm{ReLU}}(x)| \leq \varepsilon$,

(ii) *and* $\mathcal{P}(\phi) \leq \frac{2}{11}10^5 \max\{1, K^4\}R^2 d^{10} \varepsilon^{-2}$.

Proof. Let $\mathcal{F} = \mathcal{F}_K^{\mathrm{Lip,max}}$ be the K-Lipschitz-maximum catalog and suppose $L = (L_f)_{f\in\mathcal{F}}$ and $B = (B_f)_{f\in\mathcal{F}}$ are defined as for item (ii) in Example 4.3. Further, let $\xi_d \in \mathcal{C}_\mathcal{F}$, $d \in \mathbb{N}$, be given by $\xi_d = [(\mathrm{id}_{\mathbb{R}^d}, 0, (g_1, \dots, g_d)), (\mathrm{id}_{\mathbb{R}^d}, 0, \max_d)]$. Then $\mathcal{R}_C^{\xi_d} = f_d$, $\mathcal{D}_C(\xi_d) = 2$, $\mathcal{W}_C(\xi_d) = d$, and $\mathcal{T}_L(\xi_d) = \max\{1, K^2\}$. Moreover, $\mathbb{D}_B^{\xi_d} = \mathbb{B}_B^{\xi_d,1} = [-R, R]^d$ because $\mathbb{B}_B^{\xi_d,2} = B_{\max_d} = \mathbb{R}^d$. We conclude with Example 4.3, Remark 5.5, and Corollary 5.8. □

Note that this time, the approximating networks have a depth that grows linearly in the dimension since the network realizing the maximum function $\max_d$ has depth d. But the width grows as well due to the parallelized shallow networks approximating the Lipschitz functions.

The functions in the previous two propositions were approximated on bounded domains. But if one is willing to pay a slightly higher approximation cost, one can also approximate the family of functions from, e.g., Proposition 6.1 on the entire space without curse of dimensionality.

Proposition 6.3. *Let $q \in (1, \infty)$, $K \in (0, \infty)$ and suppose $g_d \colon \mathbb{R} \to \mathbb{R}$, $d \in \mathbb{N}$, are K-Lipschitz continuous on $\mathbb{R}$ with $|g_d(0)| \leq K$. Define $f_d \colon \mathbb{R}^d \to \mathbb{R}$ by $f_d(x) = \sum_{k=1}^d g_k(x_k)$. Then for all $d \in \mathbb{N}$ and $\varepsilon \in (0, \min\{1, 6K\}]$, there exists $\phi \in \mathcal{N}$ with $\mathcal{R}_{\mathrm{ReLU}}^\phi \in C(\mathbb{R}^d, \mathbb{R})$ such that*

(i) $\sup_{x\in\mathbb{R}^d}(1 + \|x\|^q)^{-1}|f_d(x) - \mathcal{R}_{\mathrm{ReLU}}^\phi(x)| \leq \varepsilon$, *and*

(ii) $\mathcal{P}(\phi) \leq \frac{3}{5}10^4 4^{r(3q+5)} \max\{1, K^{6r(q+1)}\}|d^{3r(q+1)+3}\varepsilon^{-2r}$ *for* $r = \frac{q}{q-1}$.

Proof. Let $\mathcal{F} = \mathcal{F}_K^{\mathrm{Lip}}$ be the K-Lipschitz catalog and suppose $L = (L_f)_{f\in\mathcal{F}}$ and $B = (B_f)_{f\in\mathcal{F}}$ are defined as for item (i) in Example 4.3. Let $V_d \in \mathbb{R}^{1\times d}$, $d \in \mathbb{N}$, be the matrix $V_d = (1 \ \cdots \ 1)$ with all entries 1 and let $\xi_d \in \mathcal{C}_\mathcal{F}$, $d \in \mathbb{N}$, be given by $\xi_d = [(\mathrm{id}_{\mathbb{R}^d}, 0, (g_1, \dots, g_d)), (V_d, 0, \mathrm{id}_\mathbb{R})]$. Then $\mathcal{R}_C^{\xi_d} = f_d$, $\mathcal{B}_C(\xi_d) = 1$, $\mathcal{D}_C(\xi_d) = 2$, $\mathcal{W}_C(\xi_d) = d$, and $\mathcal{T}_L(\xi_d) \leq d\max\{1, K^2\}$. Moreover, $\mathbb{D}_B^{\xi_d} = \mathbb{B}_B^{\xi_d,1} = \mathbb{R}^d$ because $\mathbb{B}_B^{\xi_d,2} = B_{\mathrm{id}_\mathbb{R}} = \mathbb{R}$. Hence, Example 4.3, Remark 5.5, and Corollary 5.7 imply Proposition 6.3. □

A statement analogue to Proposition 6.2 can be shown the same way. In the following proposition, we replace the sum by a product. Unfortunately, we cannot establish the approximation on an arbitrarily large domain since the Lipschitz constant of the product function on $[-r, r]^2$ is large for large r.

Proposition 6.4. *Let $K \in (0, \infty)$ and assume $g_d \colon \mathbb{R} \to \mathbb{R}$, $d \in \mathbb{N}$, are K-Lipschitz continuous with $|g_d(0)| \leq K$. Define $f_d \colon \mathbb{R}^d \to \mathbb{R}$ by $f_d(x) = \prod_{k=1}^d g_k(x_k)$ and set $R = 1/\sqrt{32}(K+1)$. Then for all $d \in \mathbb{N}_{\geq 2}$ and $\varepsilon \in (0, 1]$, there exists $\phi \in \mathcal{N}$ with $\mathcal{R}_{\mathrm{ReLU}}^\phi \in C(\mathbb{R}^d, \mathbb{R})$ such that*

(i) $\sup_{x\in[-R,R]^d}|f_d(x) - \mathcal{R}_{\mathrm{ReLU}}^\phi(x)| \leq \varepsilon$, *and*

(ii) $\mathcal{P}(\phi) \leq \frac{1}{4}10^5 \max\{1, K^4\}d^7\varepsilon^{-2}$.

Proof. Let $\mathcal{F} = \mathcal{F}_K^{\mathrm{Lip,prod}}$ be the K-Lipschitz-product catalog and suppose $L = (L_f)_{f\in\mathcal{F}}$ and $B = (B_f)_{f\in\mathcal{F}}$ are defined as for item (iii) in Example 4.3 (with $r = 1/\sqrt{32}$, $\delta = 1$, and $M = 23$). Further, let $\xi_d \in \mathcal{C}_\mathcal{F}$, $d \in \mathbb{N}_{\geq 2}$, be given by

$$\xi_d = \big[\big(\mathrm{id}_{\mathbb{R}^d}, 0, (g_1, \dots, g_d)\big), \big(\mathrm{id}_{\mathbb{R}^d}, 0, (\mathrm{pr}, \mathrm{id}_\mathbb{R}, \dots, \mathrm{id}_\mathbb{R})\big), \\ \big(\mathrm{id}_{\mathbb{R}^{d-1}}, 0, (\mathrm{pr}, \mathrm{id}_\mathbb{R}, \dots, \mathrm{id}_\mathbb{R})\big), \dots, \big(\mathrm{id}_{\mathbb{R}^3}, 0, (\mathrm{pr}, \mathrm{id}_\mathbb{R})\big), \big(\mathrm{id}_{\mathbb{R}^2}, 0, \mathrm{pr}\big)\big].$$

Then $\mathcal{R}_C^{\xi_d} = f_d$, $\mathcal{D}_C(\xi_d) = d$, $\mathcal{W}_C(\xi_d) = d$, and $\mathcal{T}_L(\xi_d) = \max\{1, K^2\}$. Moreover, $\mathbb{B}_B^{\xi_d,1} = [-R, R]^d$ and $\mathbb{B}_B^{\xi_d,n} = [-r, r]^2 \times \mathbb{R}^{d-n}$ for all $n \in \{2, \dots, d\}$. Hence, the fact that for all $n \in \{1, \dots, d-1\}$, $x \in [-R, R]^d$ we have $|\prod_{k=1}^n g_k(x_k)| \leq (K+1)^n R^n \leq r$ and $|g_{n+1}(x_{n+1})| \leq (K+1)R = r$ ensures $\mathbb{D}_B^{\xi_d} = [-R, R]^d$. Thus, Proposition 6.4 follows from Example 4.3, Remark 5.5, and Corollary 5.8. □

Since on large hypercubes, the quantity $\mathcal{T}_L(\xi_d)$ starts to grow exponentially in the dimension, the approximators in the proof of Proposition 6.4 can only be built on a small hypercube. But in the specific case, where all Lipschitz functions g_d in Proposition 6.4 are the identity, it has been shown that the d-dimensional product can be approximated without curse of dimensionality on arbitrarily large hypercubes; see Schwab and Zech [18, Proposition 3.3]). Applying the log-modification of our theory, we can recover this result:

Proposition 6.5. *Consider the functions $f_d\colon \mathbb{R}^d \to \mathbb{R}$, $d \in \mathbb{N}$, given by $f_d(x) = \prod_{k=1}^d x_k$, and let Let $R \in [1,\infty)$. Then for all $d \in \mathbb{N}_{\geq 2}$ and $\varepsilon \in (0, \min\{\frac{1}{2}, \frac{1}{R}\}]$, there exists $\phi \in \mathcal{N}$ with $\mathcal{R}^\phi_{\mathrm{ReLU}} \in C(\mathbb{R}^d, \mathbb{R})$ such that*

(i) $\sup_{x\in[-R,R]^d} |f_d(x) - \mathcal{R}^\phi_{\mathrm{ReLU}}(x)| \leq \varepsilon$, *and*

(ii) $\mathcal{P}(\phi) \leq \frac{3}{5}10^9 |\max\{1, \log_2(R)\}|^2 |\log_2(d)|^2 d^{10} |\log_2(\varepsilon^{-1})|^2$.

Proof. Fix $d \in \mathbb{N}_{\geq 2}$, let $\mathcal{F} = \mathcal{F}^{\mathrm{prod}}$ be the product catalog, and suppose $L = (L_f)_{f\in\mathcal{F}}$ and $B = (B_f)_{f\in\mathcal{F}}$ are defined as in Example 5.11 (with $r = R^d$, $\delta = \min\{\frac{1}{2}, \frac{1}{R}\}$, and $M = 1240d$). Further, let $\xi_d \in \mathcal{C}_\mathcal{F}$ be given by

$$\xi_d = \big[\big(\mathrm{id}_{\mathbb{R}^d}, 0, (\mathrm{pr}, \mathrm{id}_\mathbb{R}, \dots, \mathrm{id}_\mathbb{R})\big), \big(\mathrm{id}_{\mathbb{R}^{d-1}}, 0, (\mathrm{pr}, \mathrm{id}_\mathbb{R}, \dots, \mathrm{id}_\mathbb{R})\big), \dots, \big(\mathrm{id}_{\mathbb{R}^3}, 0, (\mathrm{pr}, \mathrm{id}_\mathbb{R})\big), \big(\mathrm{id}_{\mathbb{R}^2}, 0, \mathrm{pr}\big)\big].$$

Then $\mathcal{R}^{\xi_d}_C = f_d$, $\mathcal{D}_C(\xi_d) = d-1$, $\mathcal{W}_C(\xi_d) = d$, and $\mathcal{T}_L(\xi_d) = 32^{d/2}R^{d^2}$. Moreover, $\mathbb{B}^{\xi_d,n}_B = [-R^d, R^d]^2 \times \mathbb{R}^{d-n}$ for all $n \in \{1, \dots, d-1\}$. Hence, the fact that for all $n \in \{1, \dots, d-1\}$, $x \in [-R,R]^d$ we have $|\prod_{k=1}^n x_k| \leq R^d$ ensures $[-R,R]^d \subseteq \mathbb{D}^{\xi_d}_B$. We can conclude with Corollary 5.10, Remark 5.5, and Example 5.11 using

$$\log_2\big(\sqrt{d}(d-1)32^{d/2}R^{d^2}\varepsilon^{-1}\big) \leq \tfrac{23}{8}\max\{1, \log_2(R)\}d^2\log_2(d)\log_2(\varepsilon^{-1}).$$

□

Remark 6.6. If $R = 1$ in Proposition 6.5, we could actually do better and obtain

$$\mathcal{P}(\phi) \leq \tfrac{2}{3}10^9 |\log_2(d)|^2 d^6 |\log_2(\varepsilon^{-1})|^2$$

by taking $M = 1039$ and using

$$\log_2\big(\sqrt{d}(d-1)32^{d/2}\varepsilon^{-1}\big) \leq \tfrac{15}{4}d\log_2(d)\log_2(\varepsilon^{-1}).$$

Acknowledgements

We are grateful to Philippe von Wurstemberger for helpful comments and suggestions. This work has been partially supported by Swiss National Science Foundation research grant 200020 175699.

References

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

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

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

[4] Eldan, R., and Shamir, O. The power of depth for feedforward neural networks. In *29th Annual Conference on Learning Theory* (2016), V. Feldman, A. Rakhlin, and O. Shamir, Eds., vol. 49 of *Proceedings of Machine Learning Research*, PMLR, pp. 907–940.

[5] Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2016.

[6] Grohs, P., Hornung, F., Jentzen, A., and Von Wurstemberger, P. A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of Black–Scholes partial differential equations. *Accepted in Memoirs of the American Mathematical Society, arXiv:1809.02362* (2018), 124 pages.

[7] Grohs, P., Hornung, F., Jentzen, A., and Zimmermann, P. Space-time error estimates for deep neural network approximations for differential equations. *arXiv:1908.03833* (2019), 86 pages.

[8] Grohs, P., Perekrestenko, D., Elbrächter, D., and Bölcskei, H. Deep neural network approximation theory. *arXiv:1901.02220* (2019), 60 pages.

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

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

[11] Hutzenthaler, M., Jentzen, A., Kruse, T., and Nguyen, T. A. A proof that rectified deep neural networks overcome the curse of dimensionality in the numerical approximation of semilinear heat equations. *arXiv:1901.10854* (2019), 29 pages.

[12] Jentzen, A., Salimova, D., and Welti, T. A proof that deep artificial neural networks overcome the curse of dimensionality in the numerical approximation of Kolmogorov partial differential equations with constant diffusion and nonlinear drift coefficients. *arXiv:1809.07321* (2018), 48 pages.

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

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

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

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

[17] Petersen, P., and Voigtlaender, F. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks 108* (2018), 296–330.

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

[19] Voigtlaender, F., and Petersen, P. Approximation in $L^p(\mu)$ with deep ReLU neural networks. *arXiv:1904.04789* (2019), 4 pages.

[20] Yarotsky, D. Error bounds for approximations with deep ReLU networks. *Neural Networks 94* (2017), 103–114.