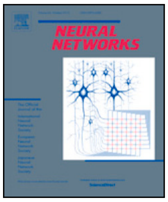
Neural Networks 180 (2024) 106702

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



Full Length Article

Near-optimal deep neural network approximation for Korobov functions with respect to L^p and H^1 norms

Yahong Yang [a,*], Yulong Lu [b]

[a] Department of Mathematics, The Pennsylvania State University, McAllister Building, Pollock Rd, State College, 16802, PA, USA
[b] School of Mathematics, University of Minnesota, Minneapolis, 206 Church St. SE, Minneapolis, 55455, MN, USA

ARTICLE INFO

Keywords:
Deep neural networks
Korobov spaces
Sobolev spaces
Optimal approximation rates

ABSTRACT

This paper derives the optimal rate of approximation for Korobov functions with deep neural networks in the high dimensional hypercube with respect to L^p-norms and H^1-norm. Our approximation bounds are non-asymptotic in both the width and depth of the networks. The obtained approximation rates demonstrate a remarkable *super-convergence* feature, improving the existing convergence rates of neural networks that are continuous function approximators. Finally, using a VC-dimension argument, we show that the established rates are near-optimal.

1. Introduction

Deep neural networks (DNNs) (Arora et al., 2016; Glorot et al., 2011) have become increasingly popular in scientific and engineering applications, including image classification (He et al., 2015; Krizhevsky et al., 2017), regularization (Czarnecki et al., 2017), and dynamic programming (Finlay et al., 2018; Werbos, 1992). In those applications, deep neural networks are often used to approximate various objects of interest, ranging from functions, functionals, to operators. Establishing quantitative universal approximation theorems of deep neural networks is an important step towards understanding their capabilities and limitations in practical applications.

Universal approximation properties of neural networks have been rigorously proved for continuous functions after the 1980s (Barron, 1993; Cybenko, 1989; Hornik et al., 1989). After that, a growing amount of literature contributed to proving quantitative approximation rates of DNNs with ReLU and square ReLU activation functions for functions with various regularity assumptions, including Besov functions (Suzuki, 2018), Sobolev functions (Gühring et al., 2020; Opschoor et al., 2020; Siegel, 2022; Yang & He, 2024; Yang, Wu, et al., 2023; Yang, Yang, & Xiang, 2023), and k-differentiable, Hölder functions. Hon and Yang (2022), Mhaskar (1996), Pinkus (1999), Shen et al. (2022), Yarotsky (2017) and holomorphic functions (Adcock et al., 2024; Opschoor et al., 2022). However, the approximation rates of DNNs in these regularity-based functions often suffer from the curse of dimensionality (CoD). For instance, the approximation rate of DNNs in Sobolev spaces $W^{n,p}([0,1]^d)$ with respect to the $W^{m,p}$ for $m < n$, $1 \le p \le \infty$, and $m, n \in \mathbb{N}$ is $\mathcal{O}\left(M^{-\frac{2(m-n)}{d}}\right)$ (up to logarithmic factors), where M is the number of parameters of the network. Notice that the rate decelerates as the d increases. The convergence rate can be substantially improved if the target function has additional low-complexity structure. Barron (1993) functions, holomorphic functions (Opschoor et al., 2022) and Korobov (1959) functions are three representative classes of functions of this kind. In fact, it has been shown that the approximation of Barron functions with shallow networks achieves a dimension-free rate of $\mathcal{O}(N^{-1/2})$ (Barron, 1993; E et al., 2022; Klusowski & Barron, 2018; Lu, Lu, & Wang, 2021; Siegel & Xu, 2022). The approximation rate can be further improved when it comes to DNN-approximation of Holomorphic functions (Opschoor et al., 2022). The work of Opschoor et al. (2022) establishes a rate of $\mathcal{O}\left(\exp\left(-bN^{1/(d+1)}\right)\right)$ in $W^{1,\infty}\left([-1,1]^d\right)$ for holomorphic functions in d dimensions, where $b > 0$ depends on the size of the region of holomorphy, and N is the size of the DNN. Approximation rates of Korobov functions with DNNs with respect to the L^∞-norm have recently been studied in Blanchard and Bennouna (2021), Mao and Zhou (2022), Montanelli and Du (2019) by leveraging tools from sparse grid approximation (Bungartz & Griebel, 2004) to overcome the CoD. In this work, we further contribute to the study along the same line and establish improved rates of convergence for Koborov functions with DNNs with respect to both L^p-norm and H^1-norm.

Let us start by giving a description of the Sobolev and Koborov function spaces. The definition of Sobolev spaces is shown as follows:

[*] Corresponding author.
E-mail address: yxy5498@psu.edu (Y. Yang).

https://doi.org/10.1016/j.neunet.2024.106702
Received 1 January 2024; Received in revised form 23 July 2024; Accepted 2 September 2024
Available online 6 September 2024

Definition 1 (*Sobolev Space* Evans, 2022)**.** Let Ω be $[0,1]^d$ and let D be the operator of the weak derivative of a single variable function and $D^{\boldsymbol{\alpha}} = D_1^{\alpha_1} D_2^{\alpha_2} \ldots D_d^{\alpha_d}$ be the partial derivative where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_d]^T$ and D_i is the derivative in the ith variable. Let $n \in \mathbb{N}$ and $1 \leq p \leq \infty$. Then we define Sobolev spaces

$$W^{n,p}(\Omega) := \left\{ f \in L^p(\Omega) : D^{\boldsymbol{\alpha}} f \in L^p(\Omega) \text{ for all } \boldsymbol{\alpha} \in \mathbb{N}^d \text{ with } |\boldsymbol{\alpha}| \leq n \right\}$$

with a norm

$$\|f\|_{W^{n,p}(\Omega)} := \left(\sum_{0 \leq |\boldsymbol{\alpha}| \leq n} \|D^{\boldsymbol{\alpha}} f\|^p_{L^p(\Omega)} \right)^{1/p}$$

if $p < \infty$, and $\|f\|_{W^{n,\infty}(\Omega)} := \max_{0 \leq |\boldsymbol{\alpha}| \leq n} \|D^{\boldsymbol{\alpha}} f\|_{L^\infty(\Omega)}$. Furthermore, for $\boldsymbol{f} = (f_1, \ldots, f_d)$, $\boldsymbol{f} \in W^{1,\infty}(\Omega, \mathbb{R}^d)$ if and only if $f_i \in W^{1,\infty}(\Omega)$ for each $i = 1, 2, \ldots, d$ and

$$\|\boldsymbol{f}\|_{W^{1,\infty}(\Omega,\mathbb{R}^d)} := \max_{i=1,\ldots,d} \{\|f_i\|_{W^{1,\infty}(\Omega)}\}.$$

When $p = 2$, denote $W^{n,2}(\Omega)$ as $H^n(\Omega)$ for $n \in \mathbb{N}_+$.

Definition 2 (*Korobov Space* Bungartz & Griebel, 2004; Korobov, 1959, 1963)**.** For $2 \leq p \leq +\infty$, the Korobov spaces $X^{2,p}(\Omega)$ is defined as

$$X^{2,p}(\Omega) = \left\{ f \in L^p(\Omega) \mid f|_{\partial\Omega} = 0, D^{\boldsymbol{k}} f \in L^p(\Omega), |\boldsymbol{k}|_\infty \leq 2 \right\}$$

with $|\boldsymbol{k}|_\infty = \max_{1 \leq j \leq d} k_j$ and the norm

$$\|f\|_{X^{2,p}(\Omega)} := \left(\sum_{0 \leq |\boldsymbol{k}|_\infty \leq 2} \left\| D^{\boldsymbol{k}} f \right\|^p_{L^p(\Omega)} \right)^{1/p}.$$

We also define the seminorm

$$|f|_{2,p} := \left\| \frac{\partial^{2d} f}{\partial x_1^2 \cdots \partial x_d^2} \right\|_{L^p(\Omega)} \tag{1}$$

Note the clear difference between Korobov space $X^{2,p}$ and the Sobolev space $W^{2,p}$: functions in the Korobov space $X^{2,p}$ have L^p-weak mixed-derivatives of up to $2d$-th order while functions in the Sobolev space $W^{2,p}$ only allow to L^p-weak derivatives up to the second order. Conversely, functions in $X^{2,p}$ demonstrate significantly lower regularity compared to those in $W^{2d,p}$ for $d > 1$. This discrepancy arises from the fact that functions in $X^{2,p}$ are only twice-differentiable in individual directions.

The neural network-approximation of Korobov functions has been studied recently in Blanchard and Bennouna (2021), Mao and Zhou (2022), Montanelli and Du (2019), Suzuki (2018). In Montanelli and Du (2019), they established an L^∞-approximation error by ReLU-DNNs for functions in $X^{2,\infty}$ with the error bound $\mathcal{O}(M^{-2})$ where M is the number of network parameters. In Blanchard and Bennouna (2021), they proved a similar convergence rate for shallow and deep networks with smooth activation function and showed that their rate is near-optimal in the sense that any continuous function approximator (DeVore et al., 1989) has a lower bound which matches up to a logarithmic factor with the established upper bound. In Suzuki (2018), the authors obtained a similar rate for mixed-Besov spaces and mixed-Sobolev spaces, which are the Korobov spaces. In Mao and Zhou (2022), the authors considered the approximation of $X^{2,p}$ using deep convolutional neural networks and proved an L^p-error bound of the form $\mathcal{O}(M^{-2+\frac{1}{p}})$. When $p = \infty$, the order is the same as that in Blanchard and Bennouna (2021), Montanelli and Du (2019), Suzuki (2018).

While significant progress has been achieved regarding the approximation of Korobov functions with DNNs, several questions remain open. Among them, the first fundamental question is to determine the optimal DNN-approximation rate of Korobov functions beyond the realm of continuous function approximators. In other words, it remains to prove whether it is possible to get a better rate for approximating Korobov functions with discontinuous function approximators. In the context introduced by DeVore et al. (1989), the term "continuous function approximators" for the approximation of neural networks means utilizing a fixed-structure neural network to approximate functions in target spaces. This process can be conceptualized as finding a mapping from the target space to the parameters in neural networks. If this mapping is continuous, the approximator of the neural network is referred to as a continuous function approximator. We formalize the mathematical definition of continuous function approximators as follows:

Definition 3 (Blanchard & Bennouna, 2021; DeVore et al., 1989)**.** Consider a subset X of a Banach space, a set of neural networks with N parameters, and an approximation scheme $G : X \to \mathbb{R}^N$ that, given an input $f \in X$, gives as output the parameters $\boldsymbol{\theta}_f = G(f)$ of the neural network approximating f. If G is continuous, then we call it a *continuous function approximator*.

Furthermore, the role of depth in the previous DNN approximation results was not carefully examined. In fact, those earlier results only showed approximation results for DNNs with either $\mathcal{O}(1)$ or $\mathcal{O}(\log(1/\epsilon))$ number of layers and with sufficient number of neurons can achieve an ϵ-accuracy. However, it remained unclear whether a diminutive approximation error could be realized by concurrently increasing both the depth and width of the network in an arbitrary fashion. Lastly, the earlier results focused on approximation error in the L^p-norm, leaving the quantification of approximation error in the Sobolev norm unexplored. As a matter of fact, Sobolev training (Czarnecki et al., 2017; Son et al., 2021; Vlassis & Sun, 2021) of DNNs has had a significant impact on scientific and engineering fields, including solving partial differential equations (De Ryck & Mishra, 2022; E et al., 2017; Lagaris et al., 1998; Raissi et al., 2019), operator learning (Liu et al., 2022; Lu, Jin, et al., 2021), network compression (Sau & Balasubramanian, 2016), distillation (Hinton et al., 2015; Rusu et al., 2015), regularization (Czarnecki et al., 2017), and dynamic programming (Finlay et al., 2018; Werbos, 1992), etc. In addition, understanding the DNN-approximation rate for Korobov functions w.r.t the Sobolev norm can benefit in theoretical understanding of neural approximation of the solution of many-body electronic Schrödinger problem as it has been shown that the ground-state of the electronic Schrödinger problem belongs to the Korobov space (Yserentant, 2004).

1.1. Contribution of the paper

We highlight the contributions of the present paper as follows.

- We first establish that a ReLU-DNN with depth $\mathcal{O}(L(\log_2 L)^{3d})$ and width $\mathcal{O}(N(\log_2 N)^{3d})$ can approximate $f \in X^{2,\infty}$ with an H^1-error of the order $\mathcal{O}(N^{-1}L^{-1})$ (see Theorem 1) and an L^p-error of the order $\mathcal{O}(N^{-2}L^{-2})$ (see Corollary 1). Notably, these outcomes align with earlier findings by Blanchard and Bennouna (2021) in the realm of continuous function approximators (DeVore et al., 1989). However, our results enhance their results by accommodating arbitrary choices of depth and width, thereby enhancing the applicability and flexibility of the established approximations.
- We next extend the study of DNN approximation of Korobov functions to the realm of discontinuous function approximators. More precisely, by adapting the bit-extraction technique (Bartlett et al., 2019, 1998) we improved the aforementioned approximation estimates to $\mathcal{O}(N^{-2}L^{-2})$ and $\mathcal{O}(N^{-4}L^{-4})$ in the context of H^1-error and L^p-error respective. See Theorems 2 and 4.
- Based on a VC-dimension argument, we show that the established bounds are near-optimal; see Theorems 3 and 5. Note that all bounds presented in the paper are non-asymptotic with respect to the network size, i.e., the approximation rate holds for all positive integers N (width) and L (depth). The results in Lu et al. (2021c), Shen et al. (2019, 2022), Yang, Yang, and Xiang (2023) are also non-asymptotic, holding for any network size. This contrasts with Bartlett et al. (2019, 1998), Cybenko (1989), Hornik (1991),

Jacot et al. (2018), Yarotsky (2018), which focus on networks with a large number of parameters or universal approximations. In Gühring et al. (2020), Gühring and Raslan (2021), Opschoor et al. (2020, 2022), Siegel (2022), network width is fixed or depth and width are correlated, with depth often being a logarithmic function of width. However, the optimality results in this paper are asymptotic since they rely on the asymptotic behavior of the VC-dimension of the neural network.

2. Preliminaries

2.1. Notations in deep neural networks

Let us summarize all basic notations used in the DNNs as follows:

1. Assume $\boldsymbol{n} \in \mathbb{N}_+^n$, then $f(\boldsymbol{n}) = \mathcal{O}(g(\boldsymbol{n}))$ means that there exists positive C independent of $\boldsymbol{n}, f, g$ such that $f(\boldsymbol{n}) \le Cg(\boldsymbol{n})$ when all entries of $\boldsymbol{n}$ go to $+\infty$.

2. Define $\sigma(x) = \max\{0, x\}$. We call the neural networks with activation function σ as σ neural networks (σ-NNs). With the abuse of notations, we define $\sigma : \mathbb{R}^d \to \mathbb{R}^d$ as $\sigma(\boldsymbol{x}) = \begin{bmatrix} \sigma(x_1) \\ \vdots \\ \sigma(x_d) \end{bmatrix}$ for any $\boldsymbol{x} = \left[x_1, \dots, x_d\right]^T \in \mathbb{R}^d$.

3. Define $L, N \in \mathbb{N}_+$, $N_0 = d$ and $N_{L+1} = 1$, $N_i \in \mathbb{N}_+$ for $i = 1, 2, \dots, L$, then a σ-NN ϕ with the width N and depth L can be described as follows:

$$\boldsymbol{x} = \tilde{\boldsymbol{h}}_0 \xrightarrow{\boldsymbol{W}_1, \boldsymbol{b}_1} \boldsymbol{h}_1 \xrightarrow{\sigma} \tilde{\boldsymbol{h}}_1 \dots \xrightarrow{\boldsymbol{W}_L, \boldsymbol{b}_L} \boldsymbol{h}_L \xrightarrow{\sigma} \tilde{\boldsymbol{h}}_L \xrightarrow{\boldsymbol{W}_{L+1}, \boldsymbol{b}_{L+1}} \phi(\boldsymbol{x}) = \boldsymbol{h}_{L+1},$$

where $\boldsymbol{W}_i \in \mathbb{R}^{N_i \times N_{i-1}}$ and $\boldsymbol{b}_i \in \mathbb{R}^{N_i}$ are the weight matrix and the bias vector in the ith linear transform in ϕ, respectively, i.e., $\boldsymbol{h}_i := \boldsymbol{W}_i \tilde{\boldsymbol{h}}_{i-1} + \boldsymbol{b}_i$, for $i = 1, \dots, L+1$ and $\tilde{\boldsymbol{h}}_i = \sigma\left(\boldsymbol{h}_i\right)$, for $i = 1, \dots, L$. In this paper, an DNN with the width N and depth L, means (a) The maximum width of this DNN for all hidden layers less than or equal to N. (b) The number of hidden layers of this DNN less than or equal to L.

2.2. Sparse-grid approximation of Korobov functions

Our approach to establishing the DNN-approximation rates for Korobov functions builds on classical approximation results of the same class of functions using sparse grids (Bungartz & Griebel, 2004). Therefore we first recall some relevant results in the sequel. For any $f \in X^{2,p}(\Omega)$, it takes the following representation:

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{l}} \sum_{\boldsymbol{i} \in \boldsymbol{i}_{\boldsymbol{l}}} v_{\boldsymbol{l},\boldsymbol{i}} \phi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x}),$$

where

$$\boldsymbol{i}_{\boldsymbol{l}} := \left\{\boldsymbol{i} \in \mathbb{N}^d : \boldsymbol{1} \le \boldsymbol{i} \le 2^{\boldsymbol{l}} - \boldsymbol{1}, i_j \text{ odd for all } 1 \le j \le d\right\}. \tag{2}$$

The basis function $\phi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x})$ is constructed using hat functions and grid points:

$$\boldsymbol{x}_{\boldsymbol{l},\boldsymbol{i}} = (x_{l_1,i_1}, \dots, x_{l_d,i_d}) := \boldsymbol{i} \odot 2^{-\boldsymbol{l}} =: \boldsymbol{i} \odot \boldsymbol{h}_{\boldsymbol{l}} = \boldsymbol{i} \odot (h_{l_1}, \dots, h_{l_d}).$$

In a piecewise linear setting, the fundamental choice for a 1D basis function is the standard hat function $\phi(x)$, defined as:

$$\phi(x) := \begin{cases} 1 - |x|, & \text{if } x \in [-1, 1] \\ 0, & \text{otherwise} \end{cases}$$

The standard hat function $\phi(x)$ can be utilized to generate any $\phi_{l_j,i_j}\left(x_j\right)$ with support $\left[x_{l_j,i_j} - h_{l_j}, x_{l_j,i_j} + h_{l_j}\right] = \left[\left(i_j - 1\right) h_{l_j}, \left(i_j + \ (1)\ h_{l_j}\right]$ through dilation and translation:

$$\phi_{l_j,i_j}\left(x_j\right) := \phi\left(\frac{x_j - i_j \cdot h_{l_j}}{h_{l_j}}\right).$$

The resulting 1D basis functions serve as inputs for the tensor product construction, yielding a suitable piecewise d-linear basis function at each grid point $\boldsymbol{x}_{\boldsymbol{l},\boldsymbol{i}}$

$$\phi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x}) := \prod_{j=1}^{d} \phi_{l_j,i_j}\left(x_j\right).$$

The following two lemmas pertain to the truncation error in the hierarchical representation of Korobov functions.

Lemma 1 (*Bungartz & Griebel, 2004, Lemma 3.3*). *Let $f \in X^{2,\infty}(\Omega)$ be given in its hierarchical representation*

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{l}} \sum_{\boldsymbol{i} \in \boldsymbol{i}_{\boldsymbol{l}}} v_{\boldsymbol{l},\boldsymbol{i}} \phi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x}).$$

Then, the following estimates for the hierarchical coefficients $v_{\boldsymbol{l},\boldsymbol{i}}$ hold:

$$\left|v_{\boldsymbol{l},\boldsymbol{i}}\right| \le 2^{-d} \cdot 2^{-|\boldsymbol{l}|_1} \cdot |f|_{2,\infty}.$$

The lemma above characterizes the decay estimates of the expansion coefficients of the Korobov space under the tensorized basis $\{\phi_{\boldsymbol{l},\boldsymbol{i}}\}$ and will play a key role in deriving the DNN-approximation rate of the main theorem.

Lemma 2 (*Bungartz & Griebel, 2004, Lemma 3.13*). *Set $f_n^{(1)}(\boldsymbol{x}) = \sum_{|\boldsymbol{l}|_1 \le n+d-1} \sum_{\boldsymbol{i} \in \boldsymbol{i}_{\boldsymbol{l}}} v_{\boldsymbol{l},\boldsymbol{i}} \phi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x})$, and for any $f \in X^{2,\infty}(\Omega)$, the approximation error satisfies*

$$\begin{aligned} \left\|f - f_n^{(1)}\right\|_{L^\infty(\Omega)} &= \mathcal{O}\left(M^{-2} \left|\log_2 M\right|^{3(d-1)}\right), \\ \left\|f - f_n^{(1)}\right\|_{H^1(\Omega)} &= \mathcal{O}\left(M^{-1} \left|\log_2 M\right|^{(d-1)}\right), \end{aligned} \tag{3}$$

where $M = \mathcal{O}(2^n n^{d-1})$.

Lemma 2 bounds the error between the sparse-grid approximation $f_n^{(1)}$ and $f \in X^{2,\infty}(\Omega)$. In the rest of the paper, we seek optimal approximations to $f_n^{(1)}$ by DNNs.

2.3. Bit-extraction technique

Proposition 1 below leverages the bit-extraction technique introduced in Bartlett et al. (2019, 1998) to represent piecewise linear functions on a fixed regular grid with M cells by a σ-NN with only $\mathcal{O}(\sqrt{M})$ parameters. Recall that the activation $\sigma = \mathrm{ReLU}$.

Proposition 1 (*Lu et al., 2021c, Proposition 4.4*). *Given any $N, L, s \in \mathbb{N}_+$ and $\xi_i \in [0, 1]$ for $i = 0, 1, \dots, N^2L^2 - 1$, there exists a σ-NN ϕ with the width $16s(N+1)\log_2(8N)$ and depth $(5L+2)\log_2(4L)$ such that*

1. $|\phi(i) - \xi_i| \le N^{-2s}L^{-2s}$ for $i = 0, 1, \dots, N^2L^2 - 1$.

2. $0 \le \phi(x) \le 1$, $x \in \mathbb{R}$.

3. Approximation in Korobov spaces with rates in continuous function approximators

In this section, we aim to establish the approximation of DNNs with an optimal rate in continuous function approximation theory. Our approximation error is dependent not only on the width N but also on the depth L of the DNNs. The result, measured by H^1 norms, is presented as follows, and the result measured by L^p norm is provided in Corollary 1.

Theorem 1. *For any $N, L \in \mathbb{N}_+$ and $f(\boldsymbol{x}) \in X^{2,\infty}(\Omega)$, there exists a continuous function approximator σ-NN $\phi(\boldsymbol{x})$ with the width $C_1 N(\log_2 N)^d$ and a depth of $C_2 L(\log_2 L)^d$ such that*

$$\|f(\boldsymbol{x}) - \phi(\boldsymbol{x})\|_{H^1(\Omega)} \le \frac{C_3}{NL} \tag{4}$$

with $\phi(\boldsymbol{x})|_{\partial\Omega} = 0$, where C_1 and C_2 are independent with N and L, and polynomially dependent on the dimension d. C_3 is dependent on $|f|_{2,\infty}$ and is independent of N and L.[1]

Remark 1. We remark that approximation rates in the main Theorems 1, 2, and the subsequent main results, including Theorem 4 and Corollary 1, are non-asymptotic in the sense that the approximation error bounds are valid for all N and L, where N is the width of the neural networks and L is the depth of the neural networks. However, we also note that the approximation lower bounds used to justify the optimality of discontinuous approximation (see Theorems 3 and 5) are asymptotic, i.e. requiring the network sizes are large enough.

Before the proof, we need to approximate the grid functions in the first.

Proposition 2. *For any $N, L \in \mathbb{N}_+$ with $|\boldsymbol{l}|_1 \leq n+d-1, \mathbf{1} \leq \boldsymbol{i} \leq 2^{\boldsymbol{l}} - \mathbf{1}$, there exists a continuous function approximator σ-NN $\hat{\phi}_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x})$ with the width $9(N+1)+4d-1$ and depth $14d(d-1)L+1$ such that*

$$\|\hat{\phi}_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x}) - \phi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x})\|_{W^{1,\infty}(\Omega)} \leq 10d^{\frac{5}{2}}(N+1)^{-7dL} \cdot 2^{|\boldsymbol{l}|_1},$$

with supp $\hat{\phi}_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x}) \subset$ supp $\phi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x})$.

Proof. For each hat function $\phi_{l_j,i_j}(x_j)$, it can be expressed as:

$$\phi_{l_j,i_j}(x_j) = \sigma\left(\frac{x_j - i_j \cdot h_{l_j}}{h_{l_j}} - 1\right) - 2\sigma\left(\frac{x_j - i_j \cdot h_{l_j}}{h_{l_j}}\right) + \sigma\left(\frac{x_j - i_j \cdot h_{l_j}}{h_{l_j}} + 1\right).$$

According to Proposition 7, there exists a σ-NN ϕ_{prod} with a width of $9(N+1)+d-1$ and depth of $14d(d-1)L$ such that $\|\phi_{\text{prod}}\|_{W^{1,\infty}([0,1]^d)} \leq 18$ and

$$\left\|\phi_{\text{prod}}(\boldsymbol{x}) - y_1 y_2 \cdots y_d\right\|_{W^{1,\infty}([0,1]^d)} \leq 10(d-1)(N+1)^{-7dL}.$$

Hence, we define $\hat{\phi}_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x}) = \phi_{\text{prod}}(\phi_{l_1,i_1}(x_1), \ldots, \phi_{l_d,i_d}(x_d))$, where $\hat{\phi}_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x})$ is a σ-NN with a width of $9(N+1)+4d-1$ and depth of $14d(d-1)L+1$. Furthermore, considering Proposition 7 and Lemma 8, we have:

$$\begin{aligned}&\|\hat{\phi}_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x}) - \phi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x})\|_{W^{1,\infty}(\Omega)} \\ &\quad = \|(\phi_{\text{prod}} - y_1 y_2 \cdots y_d) \circ (\phi_{l_1,i_1}(x_1), \ldots, \phi_{l_d,i_d}(x_d))\|_{W^{1,\infty}(\Omega)}.\end{aligned}$$

This leads to:

$$\|\hat{\phi}_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x}) - \phi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x})\|_{W^{1,\infty}(\Omega)} \leq 10d^{\frac{5}{2}}(N+1)^{-7dL} \cdot 2^{|\boldsymbol{l}|_1}.$$

Furthermore, if $\phi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x}) = 0$, there exists $\phi_{l_j,i_j}(x_j) = 0$. As per Proposition 7, we conclude $\hat{\phi}_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x}) = 0$. □

Proof of Theorem 1. Denote

$$\phi(\boldsymbol{x}) = \sum_{|\boldsymbol{l}|_1 \leq n+d-1} \sum_{\boldsymbol{i} \in \boldsymbol{i}_{\boldsymbol{l}}} v_{\boldsymbol{l},\boldsymbol{i}} \hat{\phi}_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x})$$

which can be interpreted as a σ-NN with a width of $\mathcal{O}(2^n n^{d-1} N)$ and depth of $\mathcal{O}(L)$, with the error given by

$$\begin{aligned}&\|f - \phi\|_{H^1(\Omega)} \\ &\leq C\left[M^{-1}\left|\log_2 M\right|^{(d-1)} + \sum_{|\boldsymbol{l}|_1 \leq n+d-1}\left\|\sum_{\boldsymbol{i} \in \boldsymbol{i}_{\boldsymbol{l}}}(v_{\boldsymbol{l},\boldsymbol{i}})(\hat{\phi}_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x}) - \phi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x}))\right\|_{H^1(\Omega)}\right], \quad (5)\end{aligned}$$

where the constant C is polynomially dependent on the dimension d.[2]

Due to supp $\hat{\phi}_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x}) \subset$ supp $\phi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x})$, Proposition 2, and the fact that a given $\boldsymbol{x} \in \Omega$ belongs to the support of at most one $\phi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x})$ because they have disjoint supports, we have

$$\left\|\sum_{\boldsymbol{i} \in \boldsymbol{i}_{\boldsymbol{l}}}(v_{\boldsymbol{l},\boldsymbol{i}})(\hat{\phi}_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x}) - \phi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x}))\right\|_{H^1(\Omega)} \leq 2^{-d} 2^{-|\boldsymbol{l}|_1} |f|_{2,\infty} 10d^{\frac{5}{2}}(N+1)^{-7dL}. \quad (6)$$

Since $2^{-d}\sum_{|\boldsymbol{l}|_1 \leq n+d-1} 2^{-|\boldsymbol{l}|_1} < \frac{1}{2\ln 2} \leq 1$, we have that

$$\sum_{|\boldsymbol{l}|_1 \leq n+d-1}\left\|\sum_{\boldsymbol{i} \in \boldsymbol{i}_{\boldsymbol{l}}}(v_{\boldsymbol{l},\boldsymbol{i}})(\hat{\phi}_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x}) - \phi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x}))\right\|_{H^1(\Omega)} \leq C(N+1)^{-7dL}, \quad (7)$$

where C is dependent on $|f|_{2,\infty}$ and is independent of N and L. Setting $n = \lfloor\log_2 \widetilde{N}\rfloor + \lfloor\log_2 \widetilde{L}\rfloor$, the neural network $\phi(\boldsymbol{x})$ can be viewed as a σ-NN with a depth of $\mathcal{O}(L)$ and a width of $\mathcal{O}(\widetilde{N}\widetilde{L}(\log_2 \widetilde{N} \log_2 \widetilde{L})^{d-1} N)$. It can also be regarded as the sum of a number of $\mathcal{O}(\widetilde{L}(\log_2 \widetilde{L})^{d-1}/\log_2 \widetilde{N})$ neural networks, each with a width of $\mathcal{O}(\widetilde{N}(\log_2 \widetilde{N})^d N)$ and a depth of $\mathcal{O}(L)$. Due to Proposition 9, we know that $\phi(\boldsymbol{x})$ is a σ-NN with a depth of $\mathcal{O}(L\widetilde{L}(\log_2 \widetilde{L})^{d-1}/\log_2 \widetilde{N})$ and a width of $\mathcal{O}(\widetilde{N}(\log_2 \widetilde{N})^d N)$.

Setting $N = 1$ and $L = \lfloor\log_2 \widetilde{N}\rfloor + \lfloor\log_2 \widetilde{L}\rfloor$, we have that $\phi(\boldsymbol{x})$ is a σ-NN with a width of $\mathcal{O}(\widetilde{N}(\log_2 \widetilde{N})^d)$ and a depth of $\mathcal{O}(\widetilde{L}(\log_2 \widetilde{L})^d)$. Furthermore,

$$\begin{aligned}&\sum_{|\boldsymbol{l}|_1 \leq n+d-1}\left\|\sum_{\boldsymbol{i} \in \boldsymbol{i}_{\boldsymbol{l}}}(v_{\boldsymbol{l},\boldsymbol{i}})(\hat{\phi}_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x}) - \phi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x}))\right\|_{H^1(\Omega)} \\ &\leq C(N+1)^{-7dL} \leq C2^{-7d\log_2(\widetilde{N}\widetilde{L})} \leq \frac{C}{\widetilde{N}\widetilde{L}}. \quad (8)\end{aligned}$$

Finally, due to $M \leq C\widetilde{N}\widetilde{L}(\log_2 \widetilde{N}\widetilde{L})^{d-1}$, we obtain that

$$\|f - \phi\|_{H^1(\Omega)} \leq C\left[M^{-1}\left|\log_2 M\right|^{(d-1)} + \frac{1}{\widetilde{N}\widetilde{L}}\right] \leq \frac{C}{\widetilde{N}\widetilde{L}}, \quad (9)$$

where the constant C is polynomially dependent on the dimension d. The boundary condition can be directly obtained from supp $\hat{\phi}_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x}) \subset$ supp $\phi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x})$. □

Following the same idea in the proof, we derive the following corollary, which describes the approximation of Korobov functions by deep neural networks measured by L^2 norms:

Corollary 1. *For any $N, L \in \mathbb{N}_+$ and $f(\boldsymbol{x}) \in X^{2,\infty}(\Omega)$, there exists a continuous function approximator σ-NN $\phi(\boldsymbol{x})$ with the width $C_1 N(\log_2 N)^{3d}$ and a depth of $C_2 L(\log_2 L)^{3d}$ such that*

$$\|f(\boldsymbol{x}) - \phi(\boldsymbol{x})\|_{L^p(\Omega)} \leq \frac{C_3}{N^2 L^2} \quad (10)$$

with $1 \leq p \leq \infty$ and $\phi(\boldsymbol{x})|_{\partial\Omega} = 0$, where C_1 and C_2 are independent with N and L, and depending on the dimension d at most polynomially. The constant C_3 depends on $|f|_{2,\infty}$ but does not depend on N and L.

Remark 2. Note that the number of parameters is $\mathcal{O}(N^2 L(\log_2 L)^{3d}(\log_2 N)^{6d})$, with an error of $\mathcal{O}(N^{-2}L^{-2})$. This result is consistent with the findings in Montanelli and Du (2019) when we fix N and consider depth L. Our result achieves the optimal approximation rate for continuous function approximation, as established in DeVore et al. (1989). The main improvement in our findings, compared to Blanchard and Bennouna (2021), Mao and Zhou (2022), Montanelli and Du (2019), Suzuki (2018), lies in our consideration of depth flexibility in DNNs and the establishment of the approximation rate measured by the H^1 norms.

4. Super convergence rates for Korobov functions in L^p-norms

In this section, our primary objective is to establish DNNs as function approximators within Korobov Spaces with a *super-convergence* rate, surpassing existing works. More specifically, for approximating a target function in $W^{n,p}$ measured by the norm $W^{m,p}$, we use the bit-extraction technique introduced in Barron (1993), Bartlett et al. (2019)

[1] In fact, C_1, C_2 and C_3 can be expressed by d with an explicit formula as we note in the proof of this theorem. However, the formulas may be very complicated.

[2] In this paper, we consistently employ the symbol C as a constant independent of M, N, and L, which may vary from line to line.

to approximate piecewise polynomial functions on a fixed regular grid with M cells using only $O(\sqrt{M})$ parameters. This leads to an approximation rate of $CM^{-2(n-m)/d}$ in terms of the number of parameters M, which is significantly faster than traditional methods of approximation. This phenomenon is known as the *super-convergence* of deep ReLU networks.

In this context, we focus solely on error measurement using L^p, where $1 \le p \le \infty$. In the next section, we will extend our error analysis to include Sobolev norms, specifically the H^1 norm.

Theorem 2. *For any $f \in X^{2,\infty}(\Omega)$ and $|f|_{2,\infty} \le 1$, $N, L \in \mathbb{N}_+$, there is a σ-NN $k(\boldsymbol{x})$ with $3^d(64d(N+3)(\log_2(8N))^{d+1})$ width and $(33L+2)(\log_2(4L))^{d+1}+2d$ depth such that*

$$\|k(\boldsymbol{x}) - f(\boldsymbol{x})\|_{L^p(\Omega)} \le CN^{-4}L^{-4}(\log_2 N)^{d-1}(\log_2 L)^{d-1}, \tag{11}$$

where C is the constant independent with N, L, and dependent on the $|f|_{2,\infty}$, $1 \le p \le \infty$.

Remark 3. Theorem 1 and Corollary 1 established the optimal approximation rate of Korobov spaces by continuous function approximators (Definition 3), which is also achieved previously in Blanchard and Bennouna (2021), Montanelli and Du (2019), Suzuki (2018). The new results in Theorems 2 and 4 show significant improvements over Theorem 1 and Corollary 1 due to the bit-extraction technique (Proposition 1) when the function approximators are allowed to be discontinuous, which to best of our knowledge is the first result of this kind for Korobov functions.

We split the proof of Theorem 2 into three parts. First, we prove the approximation of the Korobov function by DNNs across the entire domain Ω, excluding a small set. Then, we prove that the Korobov space consists of continuous functions. Finally, we extend the approximation to the whole domain with the help of Lemma 3 and Proposition 3.

We initially employ DNNs to approximate functions $f \in X^{2,\infty}(\Omega)$ across the entire domain Ω, excluding a small set. The domain is precisely defined as follows:

Definition 4. For any $n \ge 1$ and $|\boldsymbol{l}|_1 \le n+d-1$, let $\delta = \frac{1}{2^{n+2}}$, and

$$\Omega_{\boldsymbol{l},\delta} = \bigcup_{\boldsymbol{i}\in\boldsymbol{i}_{\boldsymbol{l}}} \Omega_{\boldsymbol{l},\boldsymbol{i},\delta}, \tag{12}$$

$$\Omega_{\boldsymbol{l},\boldsymbol{i},\delta} = \prod_{r=1}^{d}\left[\frac{2i_r-1}{2K_r}, \frac{2i_r+1}{2K_r} - \delta\cdot 1_{k<K_r-1}\right],\ K_r = 2^{l_r-1}.$$

Then we define

$$\Omega_\delta = \bigcap_{|\boldsymbol{l}|_1 \le n+d-1} \Omega_{\boldsymbol{l},\delta}.$$

Proposition 3. *For any $f \in X^{2,\infty}(\Omega)$ with $p \ge 1$ and $|f|_{2,\infty} \le 1$, $N, L \in \mathbb{N}_+$, there is a σ-NN $\tilde{k}(\boldsymbol{x})$ with $64d(N+2)(\log_2(8N))^{d+1}$ width and $(33L+2)(\log_2(4L))^{d+1}$ depth, such that*

$$\left\|\tilde{k}(\boldsymbol{x}) - f(\boldsymbol{x})\right\|_{L^\infty(\Omega_\delta)} \le CN^{-4}L^{-4}(\log_2 N)^{d-1}(\log_2 L)^{d-1}, \tag{13}$$

where C is the constant independent with N, L, and dependent on the $|f|_{2,\infty}$.

Proof. In the proof, our first step involves utilizing DNNs to approximate $\sum_{\boldsymbol{i}\in\boldsymbol{i}_{\boldsymbol{l}_*}} v_{\boldsymbol{l}_*,\boldsymbol{i}}\phi_{\boldsymbol{l}_*,\boldsymbol{i}}(\boldsymbol{x})$ for $|\boldsymbol{l}_*|_1 \le n+d-1$. Note that $\sum_{\boldsymbol{i}\in\boldsymbol{i}_{\boldsymbol{l}_*}} v_{\boldsymbol{l}_*,\boldsymbol{i}}\phi_{\boldsymbol{l}_*,\boldsymbol{i}}(\boldsymbol{x})$ can be reformulated as

$$g_{\boldsymbol{l}_*}(\boldsymbol{x}) = p_{\boldsymbol{l}_*}(\boldsymbol{x})q_{\boldsymbol{l}_*}(\boldsymbol{x}).$$

Here, $q(\boldsymbol{x})$ is a piecewise function on $[0,1]^d$ defined by

$$q_{\boldsymbol{l}_*}(\boldsymbol{x}) = v_{\boldsymbol{l}_*,\boldsymbol{i}},\ \text{for } \boldsymbol{x} \in \prod_{k=1}^{d}\left[(i_k-1)\cdot 2^{-l_k}, (i_k+1)\cdot 2^{-l_k}\right]. \tag{14}$$

Meanwhile, $p(\boldsymbol{x})$ is a piecewise-polynomial defined as

$$p_{l_k}(x) = \sum_{s=1}^{2^{k-1}} \phi\left(\frac{x_j - (2s-1)\cdot h_{l_j}}{h_{l_j}}\right)$$

$$p_{\boldsymbol{l}_*}(\boldsymbol{x}) = \prod_{k=1}^{d} p_{l_k}(x_k). \tag{15}$$

Here, $p_{l_k}(x)$ represents a deformed sawtooth function.

Let $N_r^2L_r^2 \ge K_r = 2^{l_r-1}$. By leveraging Proposition 8, we ascertain the existence of a σ-NN $\phi_r(x)$ with a width of $4N_r+5$ and a depth of $4L_r+4$ such that

$$\phi_r(x) = k, x \in \left[\frac{k}{K_r}, \frac{k+1}{K_r} - \delta\cdot 1_{k<K_r-1}\right],\ k = 1,\dots,K_r-1,$$

with $\delta \in \left(0, \frac{1}{3K^r}\right]$. Then define

$$\boldsymbol{\phi}_2(\boldsymbol{x}) = \left[\frac{\phi_1(x_1)}{K_1}, \frac{\phi_2(x_2)}{K_2}, \dots, \frac{\phi_d(x_d)}{K_d}\right]^T.$$

For each $p = 0, 1, \dots, \prod_{r=1}^d K_r - 1$, there is a bijection

$$\boldsymbol{\eta}(p) = [\eta_1, \eta_2, \dots, \eta_d] \in \prod_{r=1}^{d}\{0,\dots,K_r-1\}$$

such that $\sum_{j=1}^d \eta_j \prod_{r=1}^{j-1} K_r = p$. Set $C_{\alpha,\boldsymbol{l}_*} = 2^{-d-|\boldsymbol{l}_*|_1}|f|_{2,\infty} \ge |v_{\boldsymbol{l}_*,\boldsymbol{i}}|$ for all $\boldsymbol{i}$, and define

$$\xi_{\boldsymbol{\alpha},\boldsymbol{l}_*,\boldsymbol{i}} = \frac{v_{\boldsymbol{l}_*,\boldsymbol{i}} + C_{\alpha,\boldsymbol{l}_*}}{2C_{\alpha,\boldsymbol{l}_*}} \in [0,1]. \tag{16}$$

Based on Proposition 1, there exists a neural network $\tilde{\phi}_{\boldsymbol{\alpha}}(x)$ with a width of $16s(\tilde{N}+1)\log_2(8\tilde{N})$ and a depth of $(5\tilde{L}+2)\log_2(4\tilde{L})$ such that

$$\left|\tilde{\phi}_{\boldsymbol{\alpha}}\left(\sum_{j=1}^{d}\frac{i_j-1}{2}\prod_{r=1}^{j-1}K_r\right) - \xi_{\boldsymbol{\alpha},\boldsymbol{l}_*,\boldsymbol{i}}\right| \le \tilde{L}^{-2s}\tilde{N}^{-2s}$$

for $|\boldsymbol{l}_*|_1 \le n+d-1$ and $\boldsymbol{i} \in \boldsymbol{i}_{\boldsymbol{l}_*}$. Therefore, we define

$$\phi_{\boldsymbol{\alpha}}(\boldsymbol{x}) := 2C_{\alpha,\boldsymbol{l}_*}\tilde{\phi}_{\boldsymbol{\alpha}}\left(\sum_{j=1}^{d} x_j \prod_{r=1}^{j} K_r\right) - C_{\alpha,\boldsymbol{l}_*}. \tag{17}$$

Consequently, we find that

$$\begin{aligned}|\phi_{\boldsymbol{\alpha}}(\boldsymbol{\phi}_2(\boldsymbol{x})) - q_{\boldsymbol{l}_*}(\boldsymbol{x})| &= \left|2C_{\alpha,\boldsymbol{l}_*}\tilde{\phi}_{\boldsymbol{\alpha}}\left(\sum_{j=1}^{d}\frac{i_j-1}{2}\prod_{r=1}^{j-1}K_r\right) - C_{\alpha,\boldsymbol{l}_*} - v_{\boldsymbol{l}_*,\boldsymbol{i}}\right| \\ &\le 2C_{\alpha,\boldsymbol{l}_*}\left|\tilde{\phi}_{\boldsymbol{\alpha}}\left(\sum_{j=1}^{d}\frac{i_j-1}{2}\prod_{r=1}^{j-1}K_r\right) - \xi_{\boldsymbol{\alpha},\boldsymbol{l}_*,\boldsymbol{i}}\right| \\ &\le 2C_{\alpha,\boldsymbol{l}_*}\tilde{L}^{-2s}\tilde{N}^{-2s}\end{aligned} \tag{18}$$

for $\boldsymbol{x} \in \Omega_{\boldsymbol{l}_*,\boldsymbol{i},\delta} = \prod_{r=1}^d \left[\frac{i_r-1}{2K_r}, \frac{2i_r+1}{2K_r} - \delta\cdot 1_{k<K_r-1}\right]$.

Since there are at most 2^{n-1} elements in $\boldsymbol{i}_{\boldsymbol{l}_*}$, we set $\tilde{L}^2\tilde{N}^2 \ge 2^{n-1}$. Above all, we can let $L = \max\{\tilde{L}, L_r\} = 2^{n_1}$ and $N = \max\{\tilde{L}, N_r\} = 2^{n_2}$, where $2(n_1+n_2)$ is the smallest even number larger or equal to $n-1$. Then $\boldsymbol{\phi}_2(\boldsymbol{x})$ is a σ-NN with $4dN+5d$ width and $4L+4$ depth. $\phi_{\boldsymbol{\alpha}}(\boldsymbol{x})$ is a σ-NN with the width $16s(N+1)\log_2(8N)$ and depth $(5L+2)\log_2(4L)$. Above all, $\phi_{\boldsymbol{\alpha}}(\boldsymbol{\phi}_2(\boldsymbol{x}))$ is a σ-NN with $16sd(N+1)\log_2(8N)$ and depth $(9L+2)\log_2(4L)$. We denote this $\phi_{\boldsymbol{\alpha}}(\boldsymbol{\phi}_2(\boldsymbol{x}))$ as $s_{\boldsymbol{l}_*}(\boldsymbol{x})$, which can approximate $q_{\boldsymbol{l}_*}(\boldsymbol{x})$ well on $\Omega_{\boldsymbol{l}_*,\delta}$. Set $\delta = \frac{1}{2^{n+2}}$, we also find $s_{\boldsymbol{l}_*}(\boldsymbol{x})$, which can approximate $q_{\boldsymbol{l}_*}(\boldsymbol{x})$ well on $\Omega_{\boldsymbol{l}_*,\delta}$ for $|\boldsymbol{l}_*|_1 \le n+d-1$.

Recall

$$\Omega_\delta = \bigcap_{|\boldsymbol{l}_*|_1 \le n+d-1} \Omega_{\boldsymbol{l}_*,\delta},$$

then $s_{\boldsymbol{l}_*}(\boldsymbol{x})$, which can approximate $q_{\boldsymbol{l}_*}(\boldsymbol{x})$ well on Ω_δ for $|\boldsymbol{l}_*|_1 \le n+d-1$.

Next, we aim to approximate $p_{\boldsymbol{l}_*}(\boldsymbol{x}) = \prod_{k=1}^d p_{l_k}(x_k)$. The proof relies on leveraging the periodicity of each $p_{l_k}(x_k)$. We first define

$$\psi_1(x) = p_{l_k}(x),\ x \in [0, 2^{-n+1+n_1}],\ \text{otherwise is0}.$$

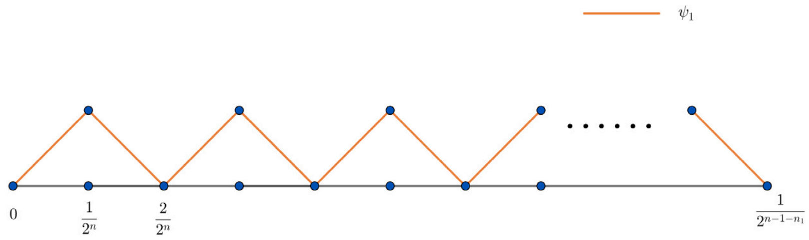


Fig. 1. The sawtooth functions ψ_1.

Then, $\psi_1(x)$ is a neural network (NN) with $4N$ width and 1 depth (see Fig. 1).

Next, we define ψ_i for $i = 2, 3, 4$ based on the symmetry and periodicity of g_i.

- ψ_2 is a function with period $\frac{2}{NL^2}$ over the interval $\left[0, \frac{1}{L^2}\right]$, where each period is represented by a hat function with a gradient of 1.
- ψ_3 is a function with period $\frac{2}{L^2}$ over the interval $\left[0, \frac{1}{L}\right]$, characterized by hat functions with a gradient of 1 for each period.
- ψ_4 is a function with period $\frac{2}{L}$ over the interval $[0, 1]$, with each period being a hat function having a gradient of 1.

Similar to ψ_1, ψ_2 is a network with a width of $4N$ and a single layer. Drawing from Proposition 9, we infer that both ψ_3 and ψ_4 are networks with a width of 7 and a depth of $L + 1$.

Finally, by Proposition 7, there exists a σ-NN $w_{l_*}(\boldsymbol{x})$ with $4(N + d + 3) + s' - 1$ width and $16s'(s' - 1)L$ depth, such that

$$\|w_{l_*}(\boldsymbol{x}) - p_{l_*}(\boldsymbol{x})\|_{L^\infty(\Omega)} \leq 10(s' - 1)(N + 1)^{-7s'L}. \tag{19}$$

Based on Proposition 6, since $|v_{l_*,i}| \leq 1$, there exists $\hat{\phi}$ with a width of $15N$ and a depth of $24L$ such that

$$\left\|\hat{\phi}(x, y) - xy\right\|_{L^\infty([-1,1]^2)} \leq 6N^{-8L}. \tag{20}$$

Therefore, we have

$$\begin{aligned}
&\|\hat{\phi}(s_{l_*}(\boldsymbol{x}), w_{l_*}(\boldsymbol{x})) - p_{l_*}(\boldsymbol{x})q_{l_*}(\boldsymbol{x})\|_{L^\infty(\Omega_\delta)} \\
\leq& \|\hat{\phi}(s_{l_*}(\boldsymbol{x}), w_{l_*}(\boldsymbol{x})) - s_{l_*}(\boldsymbol{x})w_{l_*}(\boldsymbol{x})\|_{L^\infty(\Omega_\delta)} \\
&+ \|s_{l_*}(\boldsymbol{x})w_{l_*}(\boldsymbol{x}) - s_{l_*}(\boldsymbol{x})p_{l_*}(\boldsymbol{x})\|_{L^\infty(\Omega_\delta)} \\
&+ \|s_{l_*}(\boldsymbol{x})p_{l_*}(\boldsymbol{x}) - p_{l_*}(\boldsymbol{x})q_{l_*}(\boldsymbol{x})\|_{L^\infty(\Omega_\delta)} \\
\leq& 6N^{-12L} + 20(s' - 1)(N + 1)^{-7s'L} + 4C_{\alpha,l_*}L^{-2s}N^{-2s}.
\end{aligned} \tag{21}$$

Setting $s' = s = 2$, we notice that

$$\begin{aligned}
10(s' - 1)(N + 1)^{-7s'L} &= 20(N + 1)^{-14L} \leq 20(N + 1)^{-4(L+1)} \leq 20N^{-4}L^{-4} \\
6N^{-8L} &= 6N^{4L} \cdot N^{-4(L+1)} \leq 6N^{-4}L^{-4}.
\end{aligned} \tag{22}$$

Above all, we have that there exists a σ-NN ψ_{l_*} with $64d(N + 1)\log_2(8N)$ width and $(33L + 2)\log_2(4L)$ depth such that

$$\left\|\psi_{l_*}(\boldsymbol{x}) - \sum_{i \in i_{l_*}} v_{l_*,i}\phi_{l_*,i}(\boldsymbol{x})\right\|_{L^\infty(\Omega_\delta)} \leq (26 + 4C_{\alpha,l_*})N^{-4}L^{-4}. \tag{23}$$

Similarly, we can find σ-NNs $\{\psi_l(\boldsymbol{x})\}_{|l|_1 \leq n+d-1}$ for other $\sum_{i \in i_l} v_{l,i}\phi_{l,i}(\boldsymbol{x})$ for other $\boldsymbol{l}$. Since there are at most $n^d = (2\log_2(NL) + 1)^d$ satisfied $|\boldsymbol{l}|_1 \leq n + d - 1$, we can have a σ-NN $\tilde{k}(\boldsymbol{x})$ with

$$\text{width } 32d(N + 1)\log_2(8N)(2\log_2(NL) + 1)^d \tag{24}$$

and

$$\text{depth } (33L + 2)\log_2(4L) \tag{25}$$

such that

$$\left\|\tilde{k}(\boldsymbol{x}) - f_1^{(n)}(\boldsymbol{x})\right\|_{L^\infty(\Omega_\delta)} \leq (52 + 8C_{\alpha,l_*})N^{-4}L^{-4}. \tag{26}$$

Thanks to Proposition 9, $\tilde{k}(\boldsymbol{x})$ can be expressed as

$$\frac{32d(N + 1)\log_2(8N)(2\log_2(NL) + 1)^d}{(\log_2 L)^d} \leq 64d(N + 2)(\log_2(8N))^{d+1} \tag{27}$$

width and $(33L + 2)(\log_2(4L))^{d+1}$ depth.

Finally, combining with Lemma 2 and $n = 2\log_2(NL) + 1$, we have that

$$\begin{aligned}
\|\tilde{k}(\boldsymbol{x}) - f(\boldsymbol{x})\|_{L^\infty(\Omega_\delta)} \leq& (52 + 8C_{\alpha,l_*})N^{-4}L^{-4} \\
&+ CN^{-4}L^{-4}\frac{(2\log_2(NL) + 1)^{3(d-1)}}{(2\log_2(NL) + 1)^{2(d-1)}} \\
\leq& CN^{-4}L^{-4}(\log_2 N)^{d-1}(\log_2 L)^{d-1}. \quad \square
\end{aligned} \tag{28}$$

Next, the following lemma establishes a connection between the approximation on Ω_δ and that in the whole domain.

Lemma 3 (Lu et al., 2021c; Shen et al., 2022)**.** *Given any $\varepsilon > 0, N, L, K \in \mathbb{N}^+$, and $\delta \in \left(0, \frac{1}{3K}\right]$, assume f is a continuous function in $C\left([0,1]^d\right)$ and $\widetilde{\phi}$ can be implemented by a ReLU network with width N and depth L. If*

$$|f(\boldsymbol{x}) - \widetilde{\phi}(\boldsymbol{x})| \leq \varepsilon, \quad \text{for any } \boldsymbol{x} \in \Omega_\delta,$$

then there exists a function ϕ implemented by a new ReLU network with width $3^d(N + 4)$ and depth $L + 2d$ such that

$$|f(\boldsymbol{x}) - \phi(\boldsymbol{x})| \leq \varepsilon + d \cdot \omega_f(\delta), \quad \text{for any } \boldsymbol{x} \in [0, 1]^d,$$

where

$$\omega_f(r) := \sup\left\{|f(\boldsymbol{x}) - f(\boldsymbol{y})| : \boldsymbol{x}, \boldsymbol{y} \in [0, 1]^d, \|\boldsymbol{x} - \boldsymbol{y}\|_2 \leq r\right\}, \quad \text{for any } r \geq 0. \tag{29}$$

Now, by leveraging Propositions 3, the Sobolev embedding theorem, and Lemma 3, we can derive an approximation of Korobov functions with a super-convergence rate.

Proof of Theorem 2. Based on Propositions 3, the continuity of Korobov functions, and Lemma 3, for given N, L, and d, we set δ to be sufficiently small to ensure

$$d \cdot \omega_f(\delta) \leq N^{-4}L^{-4}(\log_2 N)^{d-1}(\log_2 L)^{d-1}.$$

Then, there exists a σ-NN $k(\boldsymbol{x})$ with $3^d(64d(N + 3)(\log_2(8N))^{d+1})$ width and $2d + (33L + 2)(\log_2(4L))^{d+1}$ depth such that

$$\|k(\boldsymbol{x}) - f(\boldsymbol{x})\|_{L^\infty(\Omega)} \leq CN^{-4}L^{-4}(\log_2 N)^{d-1}(\log_2 L)^{d-1}, \tag{30}$$

where C is a constant independent of N and L, and polynomially dependent on the dimension d. Furthermore, we have

$$\|k(\boldsymbol{x}) - f(\boldsymbol{x})\|_{L^p(\Omega)} \leq CN^{-4}L^{-4}(\log_2 N)^{d-1}(\log_2 L)^{d-1}, \tag{31}$$

for any $p \in [1, \infty]$. □

The approximation rate in Theorem 2 is significantly superior to that in Corollary 1. This error outperforms the results in Blanchard and Bennouna (2021), Mao and Zhou (2022), Montanelli and Du (2019).

Furthermore, our result is nearly optimal based on the following theorem in the $X^{2,\infty}$ case.

Theorem 3. *Given any $\rho, C_1, C_2, C_3, J_0 > 0$ and $n, d \in \mathbb{N}^+$, there exist $N, L \in \mathbb{N}$ with $NL \geq J_0$ and $f \in X^{2,\infty}$ with $|f|_{2,\infty} \leq 1$, such that*

$$\inf_{\phi \in \mathcal{K}} \|\phi - f\|_{L^\infty(\Omega)} > C_3 L^{-4-\rho} N^{-4-\rho}, \tag{32}$$

where

$$\mathcal{K} := \{\sigma\text{-NNs in } \mathbb{R}^d \text{ with width } C_1 N (\log_2 N)^{d+1} \\ \text{and depth } C_2 L (\log_2 L)^{d+1}\}.$$

In order to prove this theorem, we need a definition called the Vapnik–Chervonenkis (VC) dimension (Abu-Mostafa, 1989), which describes the richness of the space. If the VC dimension of the space is large, it means that the space has greater approximating ability:

Definition 5 (*VC-dimension Abu-Mostafa, 1989*)**.** Let H denote a class of functions from $\mathcal{X}$ to $\{0, 1\}$. For any non-negative integer m, define the growth function of H as

$$\Pi_H(m) := \max_{x_1, x_2, \ldots, x_m \in \mathcal{X}} \left| \{ \left(h(x_1), h(x_2), \ldots, h(x_m) \right) : h \in H \} \right|.$$

The VC dimension of H, denoted by VCdim(H), is the largest m such that $\Pi_H(m) = 2^m$. For a class $\mathcal{G}$ of real-valued functions, define VCdim($\mathcal{G}$) := VCdim(sgn($\mathcal{G}$)), where sgn($\mathcal{G}$) := $\{\text{sgn}(f) : f \in \mathcal{G}\}$ and $\text{sgn}(x) = 1[x > 0]$.

Lemma 4 (*Bartlett et al., 2019*)**.** *For any $N, L, d \in \mathbb{N}_+$, there exists a constant $\bar{C}$ independent with N, L such that*

$$VCdim(\Phi) \leq \bar{C} N^2 L^2 \log_2 L \log_2 N, \tag{33}$$

$$\Phi := \left\{ \phi : \phi \text{ is a } \sigma - NN \text{ in } \mathbb{R}^d \text{ with width } \leq N \text{ and depth } \leq L \right\}.$$

The above lemma shows an upper bound on the VC dimension of fixed-width and fixed-depth neural networks, highlighting the limitations of such networks. This is applied in Lemma 5, which we can use to prove the optimality of Theorem 2, i.e., Theorem 3.

Lemma 5 (*Siegel, 2022*)**.** *Let $\Omega = [0, 1]^d$ and suppose that K is a translation invariant class of functions whose VC-dimension is at most n. By translation invariant we mean that $f \in K$ implies that $f(\cdot - v) \in K$ for any fixed vector $v \in \mathbb{R}^d$. Then there exists an $f \in W^{s,\infty}(\Omega)$ such that*

$$\inf_{g \in K} \|f - g\|_{L^p(\Omega)} \geq C(d, p) n^{-\frac{s}{d}} \|f\|_{W^{s,\infty}(\Omega)}.$$

Proof of Theorem 3. Define

$$\tilde{\mathcal{K}} := \{\sigma\text{-NNs in } \mathbb{R} \text{ with width } C_1 N (\log_2 N)^{d+1} \\ \text{and depth } C_2 L (\log_2 L)^{d+1}\}.$$

Due to Lemma 4, we know that

$$\text{VCdim}(\tilde{\mathcal{K}}) \leq C N^2 L^2 (\log_2 L)^{2d+3} (\log_2 N)^{2d+3}.$$

Based on Lemma 5, there exists a $\tilde{f} \in W^{2,\infty}([0, 1])$ with $\|\tilde{f}\|_{W^{2,\infty}} \leq 1$ such that

$$\inf_{g \in \mathcal{K}} \|\tilde{f} - g\|_{L^p([0,1])} \geq C(d, p) N^{-4} L^{-4} (\log_2 L)^{-4d-6} \\ \times (\log_2 N)^{-4d-6} \|\tilde{f}\|_{W^{2,\infty}([0,1])}.$$

Now we can define $f(\boldsymbol{x}) = \tilde{f}(x_1)$ which belongs to $X^{2,\infty}$. Then we know that for any $\rho, C_3, J_0 > 0$, there is an $f \in X^{2,\infty}(\Omega)$ with $|f|_{2,\infty} \leq 1$, and N, L with $NL \geq J_0$ such that

$$\begin{aligned} &\inf_{\phi \in \mathcal{K}} \|f - \phi\|_{L^\infty(\Omega)} \\ &\geq \inf_{\phi \in \mathcal{K}} \|\tilde{f}(x_1) - \phi(x_1, \ldots, x_d)\|_{L^\infty(\Omega)} \geq \inf_{\phi \in \tilde{\mathcal{K}}} \|\tilde{f}(x_1) - \phi(x_1)\|_{L^\infty([0,1])} \\ &\geq C(d, p) N^{-4} L^{-4} (\log_2 L)^{-4d-6} (\log_2 N)^{-4d-6} > C_3 L^{-4-\rho} N^{-4-\rho}. \end{aligned} \tag{34}$$

The second inequality is due to the fact that for any fixed $x_2, x_3, \ldots, x_d$, $\phi(x_1, \ldots, x_d)$ belongs to $\tilde{\mathcal{K}}$ with respect to x_1. □

5. Super convergence rates for Korobov functions in H^1-norm

In this section, we will extend our analysis from Section 4 to the case of H^1 norms. This extension ensures that our DNNs can approximate functions in Korobov spaces with minimal discrepancies in both magnitude and derivative, achieving optimality and demonstrating the *super-convergence* rate.

Theorem 4. *For any $f \in X^{2,\infty}(\Omega)$ and $|f|_{2,\infty} \leq 1$, $\|f\|_{W^{1,\infty}(\Omega)} \leq 1$, $N, L \in \mathbb{N}_+$, there is a σ-NN $k(\boldsymbol{x})$ with $2^{d+6} d(N + 2)(\log_2(8N))^{d+1}$ width and $(47L + 2)(\log_2(4L))^{d+1}$ depth such that*

$$\|f(\boldsymbol{x}) - k(\boldsymbol{x})\|_{H^1([0,1]^d)} \leq C N^{-2} L^{-2} (\log_2 N)^{d-1} (\log_2 L)^{d-1},$$

where C is the constant independent with N, L and dependent on the $|f|_{2,\infty}$.

Remark 4. The proof of Theorem 4 is different from that of Theorem 2. The reason is that the derivative in $\Omega \backslash \Omega_\delta$ is very large and cannot be estimated well by Lemma 3. Therefore, the proof of Theorem 4 can be divided into three parts. The first part involves dividing the domain Ω into several parts and finding a neural network to approximate the target function in each part as in Theorem 2. Then, we establish a partition of unity and use DNNs to approximate them (Yang, Yang, & Xiang, 2023). Finally, we combine the neural networks from the first two steps to establish a neural network that approximates the target function in the whole domain with H^1-norms.

First of all, define a sequence of subsets of Ω:

Definition 6. Given $K, d \in \mathbb{N}^+$, and for any $\boldsymbol{m} = (m_1, m_2, \ldots, m_d) \in \{1, 2\}^d$, we define $\Omega_{\boldsymbol{m}} := \prod_{j=1}^d \Omega_{m_j}$, where $\Omega_1 := \bigcup_{i=0}^{K-1} \left[\frac{i}{K}, \frac{i}{K} + \frac{3}{4K} \right]$, $\Omega_2 := \bigcup_{i=0}^{K} \left[\frac{i}{K} - \frac{1}{2K}, \frac{i}{K} + \frac{1}{4K} \right] \cap [0, 1]$.

Note that $\Omega_{\boldsymbol{1}} = \Omega_\delta$ when $K = \frac{1}{2^n}$, where n and Ω_δ are defined in Definition 4.

Next, we are going to establish neural networks on each $\Omega_{\boldsymbol{m}}$ to approximate the Korobov functions in the H^1-norm.

Proposition 4. *For any $f \in X^{2,\infty}(\Omega)$ with $p \geq 1$ and $|f|_{2,\infty} \leq 1$, $N, L \in \mathbb{N}_+$, there is a σ-NN $\tilde{k}_{\boldsymbol{m}}(\boldsymbol{x})$ for any $\boldsymbol{m} \in \{1, 2\}^d$ with $64d(N + 2)(\log_2(8N))^{d+1}$ width and $(33L + 2)(\log_2(4L))^{d+1}$ depth, such that*

$$\left\| \tilde{k}_{\boldsymbol{m}}(\boldsymbol{x}) - f(\boldsymbol{x}) \right\|_{H^1(\Omega_{\boldsymbol{m}})} \leq C N^{-2} L^{-2}. \tag{35}$$

where C is the constant independent with N, L, and dependent on the $|f|_{2,\infty}$.

Proof. The proof is similar to that of Proposition 3. We consider $\boldsymbol{m} = \boldsymbol{1}$, i.e., $\Omega_{\boldsymbol{m}_*} = \Omega_\delta$. For other $\boldsymbol{m} \in \{1, 2\}^d$, the proof can be carried out in a similar way. For any $|\boldsymbol{l}| \leq n + d - 1$, there exists a σ-NN $\psi_{\boldsymbol{l}}$ with $64d(N + 1)\log_2(8N)$ width and $(33L + 2)\log_2(4L)$ depth such that

$$\left\| \psi_{\boldsymbol{l}}(\boldsymbol{x}) - \sum_{\boldsymbol{i} \in \boldsymbol{i}_{\boldsymbol{l}}} v_{\boldsymbol{l},\boldsymbol{i}} \phi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x}) \right\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*})} \leq (26 + 4C_{\alpha,\boldsymbol{l}_*}) N^{-4} L^{-4}. \tag{36}$$

The proof follows a similar structure to that in Proposition 3. This similarity arises from the fact that $\sum_{\boldsymbol{i} \in \boldsymbol{i}_{\boldsymbol{l}}} v_{\boldsymbol{l},\boldsymbol{i}} \phi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x}) = p_{\boldsymbol{l}}(\boldsymbol{x}) q_{\boldsymbol{l}}(\boldsymbol{x})$ and $p_{\boldsymbol{l}}$ is a piece-wise constant function with a weak derivative always equal to zero. The approximation of $p_{\boldsymbol{l}}(\boldsymbol{x})$ has already been measured by the norm $W^{1,\infty}$ in Proposition 3. Due to $W^{1,\infty}(\Omega) \subset H^1(\Omega)$, we can have a σ-NN $\tilde{k}_{\boldsymbol{m}_*}(\boldsymbol{x})$ with width $32d(N+1)\log_2(8N)(2\log_2(NL)+1)^d$ and depth $(33L + 2)\log_2(4L)$ such that

$$\left\| \tilde{k}_{\boldsymbol{m}_*}(\boldsymbol{x}) - f_1^{(n)}(\boldsymbol{x}) \right\|_{H^1(\Omega_{\boldsymbol{m}_*})} \leq (52 + 8C_{\alpha,\boldsymbol{l}_*}) N^{-4} L^{-4}. \tag{37}$$

Combine with Lemma 2, we have

$$\begin{aligned} \left\| \tilde{k}_{\boldsymbol{m}_*}(\boldsymbol{x}) - f(\boldsymbol{x}) \right\|_{H^1(\Omega_{\boldsymbol{m}_*})} &\leq (52 + 8C_{\alpha,\boldsymbol{l}_*}) N^{-4} L^{-4} + C N^{-2} L^{-2} \\ &\leq C N^{-2} L^{-2}. \quad □ \end{aligned} \tag{38}$$

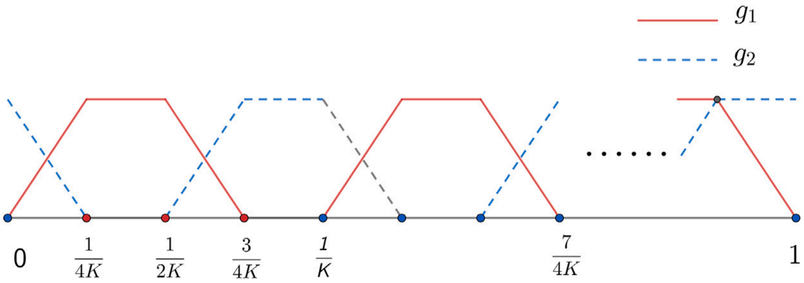


Fig. 2. The schematic diagram of g_i for $i = 1, 2$.

Then we define a partition of unity $\{g_{\boldsymbol{m}}\}_{\boldsymbol{m}\in\{1,2\}^d}$ on $[0,1]^d$ with $\operatorname{supp} g_{\boldsymbol{m}} \cap [0,1]^d \subset \Omega_{\boldsymbol{m}}$ for each $\boldsymbol{m} \in \{1,2\}^d$:

Definition 7. Given $K, d \in \mathbb{N}_+$, we define

$$g_1(x) := \begin{cases} 1, & x \in \left[\frac{i}{K} + \frac{1}{4K}, \frac{i}{K} + \frac{1}{2K}\right] \\ 0, & x \in \left[\frac{i}{K} + \frac{3}{4K}, \frac{i+1}{K}\right] \\ 4K\left(x - \frac{i}{K}\right), & x \in \left[\frac{i}{K}, \frac{i}{K} + \frac{1}{4K}\right] \\ -4K\left(x - \frac{i}{K} - \frac{3}{4K}\right), & x \in \left[\frac{i}{K} + \frac{1}{2K}, \frac{i}{K} + \frac{3}{4K}\right] \end{cases}$$

$$g_2(x) := g_1\left(x + \frac{1}{2K}\right), \tag{39}$$

for $i \in \mathbb{Z}$, (see Fig. 2). For any $\boldsymbol{m} = (m_1, m_2, \dots, m_d) \in \{1,2\}^d$, define $g_{\boldsymbol{m}}(\boldsymbol{x}) = \prod_{j=1}^d g_{m_j}(x_j)$, $\boldsymbol{x} = (x_1, x_2, \dots, x_d)$.

Then we use the following proposition to approximate $\{g_{\boldsymbol{m}}\}_{\boldsymbol{m}\in\{1,2\}^d}$ by σ-NNs and construct a sequence of σ-NNs $\{\phi_{\boldsymbol{m}}\}_{\boldsymbol{m}\in\{1,2\}^d}$. The proof shown in the Yang, Yang, and Xiang (2023).

Proposition 5 (*Yang, Yang, & Xiang, 2023*)**.** *Given any $N, L, n \in \mathbb{N}_+$ for $K = N^2L^2$, then for any*

$$\boldsymbol{m} = (m_1, m_2, \dots, m_d) \in \{1,2\}^d,$$

there is a σ-NN with the width smaller than $(9+d)(N+1)+d-1$ and depth smaller than $15d(d-1)nL$ such as

$$\|\phi_{\boldsymbol{m}}(\boldsymbol{x}) - g_{\boldsymbol{m}}(\boldsymbol{x})\|_{W^{1,\infty}([0,1]^d)} \le 50d^{\frac{5}{2}}(N+1)^{-4dnL}.$$

Now we combine $\{\hat{k}_{\boldsymbol{m}}(\boldsymbol{x})\}_{\boldsymbol{m}\in\{1,2\}^d}$ and $\{\phi_{\boldsymbol{m}}(\boldsymbol{x})\}_{\boldsymbol{m}\in\{1,2\}^d}$ in Proposition 5 to extend the approximation into the whole domain Ω. Before doing so, we require the following lemma. This lemma demonstrates that $\phi_{\boldsymbol{m}}(\boldsymbol{x})$ in Proposition 5 attains 0 to 0 behavior in the Sobolev norms.

Lemma 6 (*Yang, Yang, & Xiang, 2023*)**.** *For any $\lambda \in H^1([0,1]^d)$, we have*

$$\begin{aligned} \|\phi_{\boldsymbol{m}} \cdot \lambda\|_{H^1([0,1]^d)} &= \|\phi_{\boldsymbol{m}} \cdot \lambda\|_{H^1([0,1]^d)} \\ \|\phi_{\boldsymbol{m}} \cdot \lambda - \phi(\phi_{\boldsymbol{m}}, \lambda)\|_{H^1([0,1]^d)} &= \|\phi_{\boldsymbol{m}} \cdot \lambda - \phi(\phi_{\boldsymbol{m}}, \lambda)\|_{H^1([0,1]^d)} \end{aligned} \tag{40}$$

for any $\boldsymbol{m} \in \{1,2\}^d$, where $\phi_{\boldsymbol{m}}$ and $\Omega_{\boldsymbol{m}}$ is defined in Proposition 5 and Definition 6, and ϕ is from Proposition 6 (choosing $a = 1$ in the proposition) (see Fig. 2).

Proof of Theorem 4. Based on Propositions 3 and 4, there is a sequence of the neural network $\{\tilde{k}_{\boldsymbol{m}}(\boldsymbol{x})\}_{\boldsymbol{m}\in\{1,2\}^d}$ such that

$$\begin{aligned} &\|\tilde{k}_{\boldsymbol{m}}(\boldsymbol{x}) - f(\boldsymbol{x})\|_{H^1(\Omega_{\boldsymbol{m}})} \le CN^{-2}L^{-2}, \\ &\|\tilde{k}_{\boldsymbol{m}}(\boldsymbol{x}) - f(\boldsymbol{x})\|_{L_2(\Omega)} \le CN^{-4}L^{-4}(\log_2 N)^{d-1}(\log_2 L)^{d-1}, \end{aligned} \tag{41}$$

where C is independent with N and L, and each $\tilde{k}_{\boldsymbol{m}}(\boldsymbol{x})$ for any $\boldsymbol{m} \in \{1,2\}^d$ is a σ-NN with $64d(N+2)(\log_2(8N))^{d+1}$ width and $(33L+2)(\log_2(4L))^{d+1}$ depth. According to Proposition 5, there is a sequence of the neural network $\{\phi_{\boldsymbol{m}}(\boldsymbol{x})\}_{\boldsymbol{m}\in\{1,2\}^d}$ such that

$$\|\phi_{\boldsymbol{m}}(\boldsymbol{x}) - g_{\boldsymbol{m}}(\boldsymbol{x})\|_{W^{1,\infty}([0,1]^d)} \le 50d^{\frac{5}{2}}(N+1)^{-4dL},$$

where $\{g_{\boldsymbol{m}}\}_{\boldsymbol{m}\in\{1,2\}^d}$ is defined in Definition 7 with $\sum_{\boldsymbol{m}\in\{1,2\}^d} g_{\boldsymbol{m}}(\boldsymbol{x}) = 1$ and $\operatorname{supp} g_{\boldsymbol{m}} \cap [0,1]^d = \Omega_{\boldsymbol{m}}$. For each $\phi_{\boldsymbol{m}}$, it is a neural network with the width smaller than $(9+d)(N+1)+d-1$ and depth smaller than $15d(d-1)L$.

Due to Proposition 6, there is a neural network $\widetilde{\Phi}$ with the width $15(N+1)$ and depth $14L$ such that $\|\phi\|_{W^{1,\infty}[-1,1]^2} \le 12$ and

$$\left\|\widetilde{\Phi}(x,y) - xy\right\|_{W^{1,\infty}[-1,1]^2} \le 6(N+1)^{-7(L+1)}. \tag{42}$$

Now we define

$$k(\boldsymbol{x}) = \sum_{\boldsymbol{m}\in\{1,2\}^d} \phi(\phi_{\boldsymbol{m}}(\boldsymbol{x}), \tilde{k}_{\boldsymbol{m}}(\boldsymbol{x})). \tag{43}$$

Note that

$$\begin{aligned} \mathcal{R} :=& \|f(\boldsymbol{x}) - \phi(\boldsymbol{x})\|_{H^1([0,1]^d)} = \left\|\sum_{\boldsymbol{m}\in\{1,2\}^d} g_{\boldsymbol{m}} \cdot f(\boldsymbol{x}) - \phi(\boldsymbol{x})\right\|_{H^1([0,1]^d)} \\ \le& \left\|\sum_{\boldsymbol{m}\in\{1,2\}^d} \left[g_{\boldsymbol{m}} \cdot f(\boldsymbol{x}) - \phi_{\boldsymbol{m}}(\boldsymbol{x}) \cdot \psi_{\boldsymbol{m}}(\boldsymbol{x})\right]\right\|_{H^1([0,1]^d)} \\ &+ \left\|\sum_{\boldsymbol{m}\in\{1,2\}^d} \left[\phi_{\boldsymbol{m}}(\boldsymbol{x}) \cdot \psi_{\boldsymbol{m}}(\boldsymbol{x}) - \widetilde{\Phi}(\phi_{\boldsymbol{m}}(\boldsymbol{x}), \psi_{\boldsymbol{m}}(\boldsymbol{x}))\right]\right\|_{H^1([0,1]^d)}. \end{aligned} \tag{44}$$

As for the first part,

$$\begin{aligned} &\left\|\sum_{\boldsymbol{m}\in\{1,2\}^d} \left[g_{\boldsymbol{m}} \cdot f(\boldsymbol{x}) - \phi_{\boldsymbol{m}}(\boldsymbol{x}) \cdot \psi_{\boldsymbol{m}}(\boldsymbol{x})\right]\right\|_{H^1([0,1]^d)} \\ \le& \sum_{\boldsymbol{m}\in\{1,2\}^d} \|g_{\boldsymbol{m}} \cdot f(\boldsymbol{x}) - \phi_{\boldsymbol{m}}(\boldsymbol{x}) \cdot \psi_{\boldsymbol{m}}(\boldsymbol{x})\|_{H^1([0,1]^d)} \\ \le& \sum_{\boldsymbol{m}\in\{1,2\}^d} \left[\|(g_{\boldsymbol{m}} - \phi_{\boldsymbol{m}}(\boldsymbol{x})) \cdot f(\boldsymbol{x})\|_{H^1([0,1]^d)} + \|(f_{\boldsymbol{m}} - \psi_{\boldsymbol{m}}(\boldsymbol{x})) \cdot \phi_{\boldsymbol{m}}(\boldsymbol{x})\|_{H^1([0,1]^d)}\right] \\ =& \sum_{\boldsymbol{m}\in\{1,2\}^d} \left[\|(g_{\boldsymbol{m}} - \phi_{\boldsymbol{m}}(\boldsymbol{x})) \cdot f(\boldsymbol{x})\|_{H^1([0,1]^d)} + \|(f_{\boldsymbol{m}} - \psi_{\boldsymbol{m}}(\boldsymbol{x})) \cdot \phi_{\boldsymbol{m}}(\boldsymbol{x})\|_{H^1([0,1]^d)}\right], \end{aligned} \tag{45}$$

where the last equality is due to Lemma 6. Based on $\|f\|_{W^{1,\infty}([0,1]^d)} \le 1$, we have

$$\begin{aligned} \|(g_{\boldsymbol{m}} - \phi_{\boldsymbol{m}}(\boldsymbol{x})) \cdot f(\boldsymbol{x})\|_{H^1([0,1]^d)} &\le \|(g_{\boldsymbol{m}} - \phi_{\boldsymbol{m}}(\boldsymbol{x}))\|_{H^1([0,1]^d)} \\ &\le 50d^{\frac{5}{2}}(N+1)^{-4dnL}. \end{aligned} \tag{46}$$

And

$$\begin{aligned} &\|(f_{\boldsymbol{m}} - \psi_{\boldsymbol{m}}(\boldsymbol{x})) \cdot \phi_{\boldsymbol{m}}(\boldsymbol{x})\|_{H^1([0,1]^d)} \\ \le& \|(f_{\boldsymbol{m}} - \psi_{\boldsymbol{m}})\|_{H^1([0,1]^d)} \cdot \|\phi_{\boldsymbol{m}}\|_{L^\infty(\Omega_{\boldsymbol{m}})} + \|(f_{\boldsymbol{m}} - \psi_{\boldsymbol{m}})\|_{L^2(\Omega_{\boldsymbol{m}})} \cdot \|\phi_{\boldsymbol{m}}\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}})} \\ \le& CN^{-2}L^{-2} \cdot \left(1 + 50d^{\frac{5}{2}}\right) + CN^{-4}L^{-4} \cdot 54d^{\frac{5}{2}}N^2L^2(\log_2 N)^{d-1}(\log_2 L)^{d-1} \\ \le& CN^{-2}L^{-2}(\log_2 N)^{d-1}(\log_2 L)^{d-1}, \end{aligned} \tag{47}$$

where the second inequality is due to

$$\begin{aligned} &\|\phi_{\boldsymbol{m}}\|_{L^\infty(\Omega_{\boldsymbol{m}})} \le \|\phi_{\boldsymbol{m}}\|_{L^\infty([0,1]^d)} \le \|g_{\boldsymbol{m}}\|_{L^\infty([0,1]^d)} + \|\phi_{\boldsymbol{m}} - g_{\boldsymbol{m}}\|_{L^\infty([0,1]^d)} \le 1 + 50d^{\frac{5}{2}} \\ &\|\phi_{\boldsymbol{m}}\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}})} \le \|\phi_{\boldsymbol{m}}\|_{W^{1,\infty}([0,1]^d)} \le \|g_{\boldsymbol{m}}\|_{W^{1,\infty}([0,1]^d)} + \|\phi_{\boldsymbol{m}} - g_{\boldsymbol{m}}\|_{W^{1,\infty}([0,1]^d)} \\ &\le 4N^2L^2 + 50d^{\frac{5}{2}}. \end{aligned} \tag{48}$$

Therefore

$$\left\|\sum_{\boldsymbol{m}\in\{1,2\}^d}\left[g_{\boldsymbol{m}}\cdot f(\boldsymbol{x})-\phi_{\boldsymbol{m}}(\boldsymbol{x})\cdot\psi_{\boldsymbol{m}}(\boldsymbol{x})\right]\right\|_{W^{1,\infty}([0,1]^d)}$$
$$\le CN^{-2}L^{-2}(\log_2 N)^{d-1}(\log_2 L)^{d-1}, \tag{49}$$

due to $(N+1)^{-4dnL}\le N^{-2n}L^{-2n}$.

For the second part, due to Lemma 6, we have

$$\left\|\sum_{\boldsymbol{m}\in\{1,2\}^d}\left[\phi_{\boldsymbol{m}}(\boldsymbol{x})\cdot\psi_{\boldsymbol{m}}(\boldsymbol{x})-\widetilde{\Phi}(\phi_{\boldsymbol{m}}(\boldsymbol{x}),\psi_{\boldsymbol{m}}(\boldsymbol{x}))\right]\right\|_{H^1([0,1]^d)}$$
$$\le\sum_{\boldsymbol{m}\in\{1,2\}^d}\left\|\phi_{\boldsymbol{m}}(\boldsymbol{x})\cdot\psi_{\boldsymbol{m}}(\boldsymbol{x})-\widetilde{\Phi}(\phi_{\boldsymbol{m}}(\boldsymbol{x}),\psi_{\boldsymbol{m}}(\boldsymbol{x}))\right\|_{H^1([0,1]^d)}$$
$$=\sum_{\boldsymbol{m}\in\{1,2\}^d}\left\|\phi_{\boldsymbol{m}}(\boldsymbol{x})\cdot\psi_{\boldsymbol{m}}(\boldsymbol{x})-\widetilde{\Phi}(\phi_{\boldsymbol{m}}(\boldsymbol{x}),\psi_{\boldsymbol{m}}(\boldsymbol{x}))\right\|_{H^1(\Omega_{\boldsymbol{m}})}. \tag{50}$$

Due to Lemma 8, we have that

$$\left\|\phi_{\boldsymbol{m}}(\boldsymbol{x})\cdot\psi_{\boldsymbol{m}}(\boldsymbol{x})-\widetilde{\Phi}(\phi_{\boldsymbol{m}}(\boldsymbol{x}),\psi_{\boldsymbol{m}}(\boldsymbol{x}))\right\|_{H^1(\Omega_{\boldsymbol{m}})}$$
$$\le CN^{-2}L^{-2}(\log_2 N)^{d-1}(\log_2 L)^{d-1}. \tag{51}$$

Combining (49) and (51), we have that there is a σ-NN with $(47L+2)(\log_2(4L))^{d+1}$ depth and $2^{d+6}d(N+2)(\log_2(8N))^{d+1}$ width such that

$$\|f(\boldsymbol{x})-k(\boldsymbol{x})\|_{H^1([0,1]^d)}\le CN^{-2}L^{-2}(\log_2 N)^{d-1}(\log_2 L)^{d-1},$$

where C is the constant independent with N,L. □

Remark 5. The approximation rate for Korobov functions provided in Theorem 4 falls short of achieving the nearly optimal approximation rate observed in function spaces $W^{2d,p}$, as measured by the norm containing the first derivative (Yang, Yang, & Xiang, 2023). In the latter case, the optimal rate is $\mathcal{O}((NL)^{-\frac{4d-2}{d}})$. The limitation in achieving this optimal rate for Korobov functions is rooted in the sensitivity of these functions to derivatives. For instance, consider a finite expansion of f in Korobov spaces denoted as $f_n^{(1)}(\boldsymbol{x})=\sum_{|\boldsymbol{l}|_1\le n+d-1}\sum_{\boldsymbol{i}\in\boldsymbol{i}_{\boldsymbol{l}}}v_{\boldsymbol{l},\boldsymbol{i}}\phi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x})$. In this expansion, there exists a spline function $\phi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x})$ for which $\boldsymbol{l}=(n,1,1,1,\dots,1)$, and its partial derivative with respect to x_1 can be very large, on the order of 2^n.

The way to prove the optimality of the H^1 case is similar to Theorem 3 and combined with the following lemma:

Lemma 7 (*Yang, Yang, & Xiang, 2023, Theorem 1*). *For any $N,L,d\in\mathbb{N}_+$, there exists a constant $\bar{C}$ independent with N,L such that*

$$\mathrm{VCdim}(D\Phi)\le\bar{C}N^2L^2\log_2 L\log_2 N,$$

for

$$D\Phi:=\left\{\psi=D_i\phi:\phi\in\Phi,\ i=1,2,\dots,d\right\}, \tag{52}$$

where $\Phi:=\left\{\phi:\phi\ is\ a\ \sigma-NN\ in\ \mathbb{R}^d\ with\ width\le N\ and\ depth\ \le L\right\}$, and D_i is the weak derivative in the ith variable.

Theorem 5. *Given any $\rho,C_1,C_2,C_3,J_0>0$ and $n,d\in\mathbb{N}^+$, there exist $N,L\in\mathbb{N}$ with $NL\ge J_0$ and $f\in X^{2,\infty}$ with $|f|_{2,\infty}\le 1$, such that*

$$\inf_{\phi\in\mathcal{K}}\|\phi-f\|_{H^1(\Omega)}>C_3L^{-2-\rho}N^{-2-\rho}, \tag{53}$$

where

$$\mathcal{K}:=\{\sigma\text{-NNs in }\mathbb{R}^d\text{ with width }C_1N(\log_2 N)^{d+1}$$
$$\text{and depth }C_2L(\log_2 L)^{d+1}\}.$$

Proof. The proof is similar to Theorem 3 and combined with Lemma 7. □

Comparing the results in Theorem 1 and Corollary 1 with Theorems 2 and 4, we observe that the results in Theorems 2 and 4 are significantly better than those in Theorem 1 and Corollary 1. The constants in Theorem 1 and Corollary 1 are superior to those in Theorems 2 and 4, which exponentially depend on the dimension d. This leaves an open question for future research to explore alternative approaches for addressing the challenge of incorporating the dependence on d in the lower bounds while maintaining a *super-convergence* rate.

6. Conclusion

This paper establishes the approximation of DNNs for Korobov functions, not only in L^p norms for $2\le p\le\infty$ but also in H^1 norms, effectively avoiding the curse of dimensionality. For both types of errors, we establish a *super-convergence* rate and prove the optimality of each approximation.

In our exploration of deep neural networks for approximating Korobov functions, we note that prior work, such as Blanchard and Bennouna (2021), has focused on two-hidden layer neural networks for shallow approximations. The establishment of the potential of one-hidden layer neural networks for approximating functions in Korobov spaces is considered as future work. Moreover, in this paper, we delve into proving the optimality of our results. The proof strategy relies on the fact that the approximation rate in $X^{2,\infty}([0,1]^d)$ achieves a nearly optimal approximation rate for $W^{2,\infty}([0,1])$. However, when combining our work with the estimates provided in Mao and Zhou (2022), it becomes evident that the *super-convergence* rate for $X^{2,p}$ can only achieve $\mathcal{O}\left(N^{-4+\frac{2}{p}}L^{-4+\frac{2}{p}}\right)$ (up to logarithmic factors). Determining whether this rate is nearly optimal and establishing a proof for it remains an open question.

CRediT authorship contribution statement

Yahong Yang: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Yulong Lu:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgment

YL is supported by the National Science Foundation, United States through the award DMS-2343135.

Appendix. Preliminary results on ReLU-DNN approximation

In this section, we collect several lemmas and propositions related to DNNs. The following two propositions concern the approximation of product operators in the $W^{1,\infty}$ sense. These will be used in the proof to represent the sparse grid and combine the neural network in the whole domain, excluding small sets and the partition of unity.

Proposition 6 (Yang, Yang, & Xiang, 2023). *For any $N, L \in \mathbb{N}_+$ and $a > 0$, there is a σ-NN ϕ with the width $15N$ and depth $2L$ such that $\|\phi\|_{W^{1,\infty}((-a,a)^2)} \le 12a^2$ and*

$$\|\phi(x,y) - xy\|_{W^{1,\infty}((-a,a)^2)} \le 6a^2 N^{-L}. \tag{A.1}$$

Furthermore,

$$\phi(0,y) = \frac{\partial\phi(0,y)}{\partial y} = 0, \;\; y \in (-a,a). \tag{A.2}$$

Proposition 7 (Yang, Yang, & Xiang, 2023). *For any $N, L, s \in \mathbb{N}_+$with $s \ge 2$, there exists a σ-NN ϕ with the width $9(N+1) + s - 1$ and depth $14s(s-1)L$ such that $\|\phi\|_{W^{1,\infty}((0,1)^s)} \le 18$ and*

$$\left\|\phi(\boldsymbol{x}) - x_1 x_2 \cdots x_s\right\|_{W^{1,\infty}((0,1)^s)} \le 10(s-1)(N+1)^{-7sL}. \tag{A.3}$$

Furthermore, for any $i = 1, 2, \ldots, s$, if $x_i = 0$, we will have

$$\phi(x_1, x_2, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_s) = \frac{\partial \phi(x_1, x_2, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_s)}{\partial x_j} = 0, \tag{A.4}$$

for $i \neq j$.

In the paper, we will use a lemma concerning the composition of functions in Sobolev spaces:

Lemma 8 (Gühring et al., 2020, Corollary B.5). *Let $d, m \in \mathbb{N}_+$ and $\Omega_1 \subset \mathbb{R}^d$ and $\Omega_2 \subset \mathbb{R}^m$ both be open, bounded, and convex. Then for $\boldsymbol{f} \in W^{1,\infty}(\Omega_1, \Omega_2)$ and $g \in W^{1,\infty}(\Omega_2)$, we have*

$$\|g \circ \boldsymbol{f}\|_{W^{1,\infty}(\Omega_2)} \le \sqrt{d}m \max\{\|g\|_{L^\infty(\Omega_2)}, \|g\|_{W^{1,\infty}(\Omega_2)} \|\boldsymbol{f}\|_{W^{1,\infty}(\Omega_1, \mathbb{R}^m)}\}.$$

The following proposition, combined with the bit extraction technique, can be applied to approximate the piecewise linear function well in the whole domain, excluding a small set.

Proposition 8 (Lu et al., 2021c, Proposition 4.3). *Given any $N, L \in \mathbb{N}_+$ and $\delta \in \left(0, \frac{1}{3K}\right]$ for $K \le N^2 L^2$, there exists a σ-NN ϕ with the width $4N + 5$ and depth $4L + 4$ such that*

$$\phi(x) = k, x \in \left[\frac{k}{K}, \frac{k+1}{K} - \delta \cdot 1_{k<K-1}\right], \;\; k = 0, 1, \ldots, K-1.$$

The following lemma is used to restructure a wide neural network with a wide last layer into a deep neural network with less width, which can effectively prune the neural network and make its structure match our desired specifications.

Proposition 9 (Siegel, 2022, Proposition 1). *Given a sequence of the neural network $\{p_i\}_{i=1}^M$, and each p_i is a σ-NN from $\mathbb{R}^d \to \mathbb{R}$ with the width N and depth L_i, then $\sum_{i=1}^M p_i$ is a σ-NN with the width $N + 2d + 2$ and depth $\sum_{i=1}^M L_i$.*

References

Abu-Mostafa, Y. (1989). The Vapnik-Chervonenkis dimension: Information versus complexity in learning. *Neural Computation*, *1*(3), 312–317.

Adcock, B., Brugiapaglia, S., Dexter, N., & Moraga, S. (2024). Learning smooth functions in high dimensions: from sparse polynomials to deep neural networks. arXiv preprint arXiv:2404.03761.

Arora, R., Basu, A., Mianjy, P., & Mukherjee, A. (2016). Understanding deep neural networks with rectified linear units. arXiv preprint arXiv:1611.01491.

Barron, A. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, *39*(3), 930–945.

Bartlett, P., Harvey, N., Liaw, C., & Mehrabian, A. (2019). Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, *20*(1), 2285–2301.

Bartlett, P., Maiorov, V., & Meir, R. (1998). Almost linear VC dimension bounds for piecewise polynomial networks. *Advances in Neural Information Processing Systems*, *11*.

Blanchard, M., & Bennouna, M. A. (2021). Shallow and deep networks are near-optimal approximators of korobov functions. In *International conference on learning representations*.

Bungartz, H., & Griebel, M. (2004). Sparse grids. *Acta Numerica*, *13*, 147–269.

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

Czarnecki, W., Osindero, S., Jaderberg, M., Swirszcz, G., & Pascanu, R. (2017). Sobolev training for neural networks. *Advances in Neural Information Processing Systems*, *30*.

De Ryck, T., & Mishra, S. (2022). Error analysis for physics-informed neural networks (PINNs) approximating Kolmogorov PDEs. *Advances in Computational Mathematics*, *48*(6), 1–40.

DeVore, R., Howard, R., & Micchelli, C. (1989). Optimal nonlinear approximation. *Manuscripta Mathematica*, *63*, 469–478.

E, W., Han, J., & Jentzen, A. (2017). Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *Communications in Mathematics and Statistics*, *5*(4), 349–380.

E, W., Ma, C., & Wu, L. (2022). The Barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, *55*(1), 369–406.

Evans, L. (2022). *Partial differential equations. Volume 19*. American Mathematical Society.

Finlay, C., Calder, J., Abbasi, B., & Oberman, A. (2018). Lipschitz regularized deep neural networks generalize and are adversarially robust. arXiv preprint arXiv:1808.09540.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (pp. 315–323). JMLR Workshop and Conference Proceedings.

Gühring, I., Kutyniok, G., & Petersen, P. (2020). Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *Analysis and Applications*, *18*(05), 803–859.

Gühring, I., & Raslan, M. (2021). Approximation rates for neural networks with encodable weights in smoothness spaces. *Neural Networks*, *134*, 107–130.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026–1034).

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Hon, S., & Yang, H. (2022). Simultaneous neural network approximation for smooth functions. *Neural Networks*, *154*, 152–164.

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

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

Jacot, A., Gabriel, F., & Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, *31*.

Klusowski, J., & Barron, A. (2018). Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ_1 and ℓ_0 controls. *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, *64*(12), 7649–7656.

Korobov, N. (1959). On the approximate solution of integral equations. *Doklady Akademii Nauk SSSR*, *128*, 233–238.

Korobov, N. (1963). *Number-theoretic methods in approximate analysis*. Moscow: Fizmatgiz.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90.

Lagaris, I., Likas, A., & Fotiadis, D. (1998). Artificial neural networks for solving ordinary and partial differential equations. *IEEE Transactions on Neural Networks*, *9*(5), 987–1000.

Liu, H., Yang, H., Chen, M., Zhao, T., & Liao, W. (2022). Deep nonparametric estimation of operators between infinite dimensional spaces. arXiv preprint arXiv:2201.00217.

Lu, L., Jin, P., Pang, G., Zhang, Z., & Karniadakis, G. (2021). Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, *3*(3), 218–229.

Lu, J., Lu, Y., & Wang, M. (2021). A priori generalization analysis of the Deep Ritz method for solving high dimensional elliptic partial differential equations. In *Conference on learning theory* (pp. 3196–3241). PMLR.

Lu, J., Shen, Z., Yang, H., & Zhang, S. (2021). Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, *53*(5), 5465–5506.

Mao, T., & Zhou, D. (2022). Approximation of functions from korobov spaces by deep convolutional neural networks. *Advances in Computational Mathematics*, *48*(6), 84.

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

Montanelli, H., & Du, Q. (2019). New error bounds for deep ReLU networks using sparse grids. *SIAM Journal on Mathematics of Data Science*, *1*(1), 78–92.

Opschoor, J. A., Petersen, P. C., & Schwab, C. (2020). Deep ReLU networks and high-order finite element methods. *Analysis and Applications*, *18*(05), 715–770.

Opschoor, J., Schwab, C., & Zech, J. (2022). Exponential ReLU DNN expression of holomorphic maps in high dimension. *Constructive Approximation*, *55*(1), 537–582.

Pinkus, A. (1999). Approximation theory of the MLP model in neural networks. *Acta Numerica*, *8*, 143–195.

Raissi, M., Perdikaris, P., & Karniadakis, G. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, *378*, 686–707.

Rusu, A. A., Colmenarejo, S., Gulcehre, C., Desjardins, G., Kirkpatrick, J., Pascanu, R., Mnih, V., Kavukcuoglu, K., & Hadsell, R. (2015). Policy distillation. arXiv preprint arXiv:1511.06295.

Sau, B., & Balasubramanian, V. (2016). Deep model compression: Distilling knowledge from noisy teachers. arXiv preprint arXiv:1610.09650.

Shen, Z., Yang, H., & Zhang, S. (2019). Nonlinear approximation via compositions. *Neural Networks*, *119*, 74–84.

Shen, Z., Yang, H., & Zhang, S. (2022). Optimal approximation rate of ReLU networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées*, *157*, 101–135.

Siegel, J. (2022). Optimal approximation rates for deep ReLU neural networks on Sobolev spaces. arXiv preprint arXiv:2211.14400.

Siegel, J., & Xu, J. (2022). Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks. *Foundations of Computational Mathematics*, 1–57.

Son, H., Jang, J., Han, W., & Hwang, H. (2021). Sobolev training for the neural network solutions of pdes. arXiv preprint arXiv:2101.08932.

Suzuki, T. (2018). Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *International conference on learning representations*.

Vlassis, N., & Sun, W. (2021). Sobolev training of thermodynamic-informed neural networks for interpretable elasto-plasticity models with level set hardening. *Computer Methods in Applied Mechanics and Engineering*, *377*, Article 113695.

Werbos, P. (1992). Approximate dynamic programming for real-time control and neural modeling. *Handbook of intelligent control*. Van Nostrand.

Yang, Y., & He, J. (2024). Deeper or wider: A perspective from optimal generalization error with Sobolev loss. In *Forty-first international conference on machine learning*.

Yang, Y., Wu, Y., Yang, H., & Xiang, Y. (2023). Nearly optimal approximation rates for deep super ReLU networks on Sobolev spaces. arXiv preprint arXiv:2310.10766.

Yang, Y., Yang, H., & Xiang, Y. (2023). Nearly optimal VC-dimension and pseudo-dimension bounds for deep neural network derivatives. In *Conference on neural information processing systems*.

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

Yarotsky, D. (2018). Optimal approximation of continuous functions by very deep ReLU networks. In *Conference on learning theory* (pp. 639–649). PMLR.

Yserentant, H. (2004). On the regularity of the electronic Schrödinger equation in Hilbert spaces of mixed derivatives. *Numerische Mathematik*, *98*, 731–759.