Approximation of functionals on Korobov spaces with Fourier Functional Networks

Liu, Peilin; Liu, Yuqing; Zhou, Xiang; Zhou, Ding-Xuan

Published in:
Neural Networks

Published: 01/02/2025

Published version (DOI):
10.1016/j.neunet.2024.106922

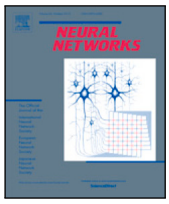
Publication details:
Liu, P., Liu, Y., Zhou, X., & Zhou, D.-X. (2025). Approximation of functionals on Korobov spaces with Fourier Functional Networks. *Neural Networks*, *182*, Article 106922. https://doi.org/10.1016/j.neunet.2024.106922

Neural Networks 182 (2025) 106922

Contents lists available at ScienceDirect

Neural Networks

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



Full Length Article

Approximation of functionals on Korobov spaces with Fourier Functional Networks

Peilin Liu [a], Yuqing Liu [b], Xiang Zhou [b,c], Ding-Xuan Zhou [a,*]

[a] School of Mathematics and Statistics, University of Sydney, Sydney, New South Wales 2006, Australia
[b] School of Data Science, City University of Hong Kong, Kowloon, Hong Kong
[c] Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong

ARTICLE INFO

Keywords:
Neural network
Approximation theory
Fourier neural operator
Convolutional neural network
Korobov space

ABSTRACT

Learning from functional data with deep neural networks has become increasingly useful, and numerous neural network architectures have been developed to tackle high-dimensional problems raised in practical domains. Despite the impressive practical achievements, theoretical foundations underpinning the ability of neural networks to learn from functional data largely remain unexplored. In this paper, we investigate the approximation capacity of a functional neural network, called Fourier Functional Network, consisting of Fourier neural operators and deep convolutional neural networks with a great reduction in parameters. We establish rates of approximating by Fourier Functional Networks nonlinear continuous functionals defined on Korobov spaces of periodic functions. Finally, our results demonstrate dimension-independent convergence rates, which overcomes the curse of dimension.

1. Introduction

Deep Learning (LeCun, Bengio, & Hinton, 2015) has achieved remarkable success in processing big data from many practical domains with its superiority in approximation, expressivity, and generalization. Along with the achievements in speech recognition, computer vision, and natural language processing, it is worth noting the recent breakthroughs in scientific research by deep learning methods such as AlphaFold (Senior et al., 2020) for predicting protein molecule structures. To address diverse challenges emerging from various research fields, neural network architectures (Anandkumar et al., 2020; Karniadakis et al., 2021; Li et al., 2020; Raissi, Perdikaris, & Karniadakis, 2019) have been developed that possess capabilities to process function-input cases with more complex data structures than speeches, images and texts. However, with their impressive performances in solving scientific problems in practice comes a paucity of theoretical understanding about how they work so well across different domains. Before introducing the latest applications and theoretical results about learning with data from function spaces by neural networks, we first recall some definitions and approximation results of neural networks defined on the Euclidean space $\mathbb{R}^d$.

Since the late 1980s, approximation properties (Barron, 1993, Cybenko, 1989; Klusowski & Barron, 2018) have been well studied with the classical shallow neural networks of form

$$f_N(x) = \sum_{k=1}^{N} c_k \sigma(a_k \cdot x + b_k)$$

with $a_k \in \mathbb{R}^d, b_k, c_k \in \mathbb{R}$ and σ an activation function from $\mathbb{R}$ to $\mathbb{R}$. The fully connected multilayer neural network (FNN) of layer L is defined by applying shallow neural networks inductively,

$$f^{(l)}(x) = \sigma\left(A^{(l)} f^{(l-1)}(x) + b^{(l)}\right), \quad l = 1, 2, \dots, L$$

where the activation function σ acts element-wise, $d_l \in \mathbb{N}$ is the number of hidden neurons (width) in lth layer, $A^{(l)}$ is a $d_l \times d_{l-1}$ matrix without special structures, $b^{(l)} \in \mathbb{R}^{d_l}$ and $f^{(0)}(x) = x$ with $d_0 = d$. The expressivity of FNNs was also well explored in Narhar Mhaskar (1993), Petersen and Voigtlaender (2018), Pinkus (1999), Yarotsky (2017).

Later, when deep learning started with the mission to reduce redundant parameters and improve computational efficiency, convolutional neural networks (CNNs) were proposed with the mechanism of weight sharing. Convolutional kernels in CNNs induced by 1-D convolutions were formally defined by Toeplitz matrices in Zhou (2020a, 2020b). For a 1-D convolutional filter $\omega = (\omega_k)_{k\in\mathbb{Z}}$ supported in $\{0, 1, \dots, s\}$ and an input $x = (x_k)_{k\in\mathbb{Z}}$ supported in $\{1, 2, \dots, d\}$, the 1-D convolution between ω and x is defined as

$$(\omega * x)_i = \sum_{k\in\mathbb{Z}} \omega_{i-k} x_k = \sum_{k=1}^{d} \omega_{i-k} x_k, \quad i \in \mathbb{Z} \tag{1}$$

* Corresponding author.
E-mail addresses: peilin.liu@sydney.edu.au (P. Liu), yuqinliu6-c@my.cityu.edu.hk (Y. Liu), xizhou@cityu.edu.hk (X. Zhou), dingxuan.zhou@sydney.edu.au (D.-X. Zhou).

https://doi.org/10.1016/j.neunet.2024.106922
Received 12 May 2024; Received in revised form 23 September 2024; Accepted 12 November 2024
Available online 20 November 2024

which by considering possibly nonzero entries of $\omega * x$, gives a $(d+s)\times d$ Toeplitz matrix T^{ω}

$$T^{\omega} := \begin{bmatrix} \omega_0 & 0 & 0 & 0 & \dots & 0 & 0 \\ \omega_1 & \omega_0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \omega_s & \omega_{s-1} & \dots & \omega_0 & \dots & 0 & 0 \\ 0 & \omega_s & \dots & \omega_1 & \ddots & \vdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \dots & \dots & 0 & \omega_s & \dots & \omega_1 & \omega_0 \\ \dots & \dots & \dots & 0 & \omega_s & \cdots & \omega_1 \\ \vdots & \dots & \dots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & \dots & 0 & \omega_s \end{bmatrix}.$$

Similarly with the fully connected neural networks, a deep convolutional neural network (DCNN) is defined by the iteration

$$f^{l}(x) = \sigma(T^{(l)} f^{(l-1)}(x) + b^{(l)}), \quad l = 1, 2, \dots, L$$

where $T^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}$ is a Toeplitz matrix, $b^{(l)} \in \mathbb{R}^{d_l}$, $d_l = d + sl$, and $f^{(0)}(x) = x$ with $d_0 = d$.

The FNNs and CNNs defined above show great superiority in tasks in speech recognition, computer vision, and natural language processing, but cannot be applied directly to the tasks of learning operators defined on function spaces with infinite dimensions.

Recently, Song, Fan, Chen, and Zhou (2023), Song, Liu, Fan, and Zhou (2023) proposed functional nets to learn functionals with certain smoothness and Shi, Yu, and Zhou (2023), Yu and Zhou (2023) showed some results in distribution regression with neural networks. Although the recent works (Song, Liu et al., 2023) achieve almost optimal rates in the respective learning tasks, the claimed network architectures declined to meet those applied in practice, e.g., Fourier Neural Operators (FNOs) and Physic-informed neural networks (PINNs). FNOs (Li et al., 2020) are inspired by the integral form of solutions to partial differential equations and extensively used in scientific computing. Convergence rates of approximating solutions to some PDEs by FNOs are studied in Kovachki, Lanthaler, and Mishra (2021), while approximating an operator with a more general smooth condition by FNOs has not been considered yet. Apart from the network architecture, the input function space also plays an important role in learning from functional data. For example, Song, Fan et al. (2023) took $L^p([-1,1]^d)$ as the input space and obtained the approximation rate $O\left(\left(\frac{\log(M)}{\log(\log(M))}\right)^{-\beta\lambda/d}\right)$ with M nonzero parameters, when the functional is defined on the unit ball of the Hölder space $C^{\beta}[-1,1]^d$ and is Lipschitz λ with $0 < \lambda \le 1$. To alleviate the effect of the curse of dimension, the Korobov spaces are considered here. Compared with the works (Mao & Zhou, 2022; Montanelli & Du, 2019) on a Korobov space of functions vanishing on the boundary of a cube, the Korobov space discussed in this paper is a reproducing kernel Hilbert space (RKHS) of periodic functions, and the capacity and the efficiency of approximation by the Korobov space are intrinsically related to the kernel functions of which polynomial cases and exponential cases are taken into consideration.

In this paper, we are interested in approximating a nonlinear continuous operator $F : H_{d,\alpha,\gamma} \to \mathbb{R}$ by a network consisting of the FNO and a DCNN, and $H_{d,\alpha,\gamma}$ is a Korobov space of periodic functions where a network with an FNO is defined. We establish a theory to show the ability of FNOs to extract features of functions from Korobov spaces. Then we construct a DCNN with multiple channels to realize the high-dimensional interpolation with a great reduction in the number of parameters from $O(d^2)$ to $O(\log_2 d)$. Finally, a convergence rate, which beats the curse of dimension, is achieved by our proposed network. The remainder of this paper is organized as follows. In Section 2, the definitions of Korobov Spaces and Fourier Functional Networks will be introduced. The main results will be established in Section 3, and the proofs of the main results are presented in Section 4.

2. Definitions

2.1. Korobov space

The Korobov space $H_{d,\alpha,\gamma}$ of one-periodic functions is a separable Hilbert space with complex-valued functions which can be specified by values on $\mathbb{T}^d := [0,1]^d$. The parameter $\alpha \ge 0$ or $\alpha = \infty$ measures the smoothness of these functions. Throughout the paper, we always assume $\alpha > 1$ in which case $H_{d,\alpha,\boldsymbol{\gamma}}$ becomes a reproducing kernel Hilbert space of periodic functions.

Let $\{a_j\}$ and $\{b_j\}$ be two positive sequences such that $1 \ge a_1 \ge a_2 \ge \cdots > 0$ and $b := \inf_{j\in\mathbb{N}} b_j > 0$. Then $\gamma = \{\gamma_j\}_{j\in\mathbb{N}}$ can be expressed in terms of the two sequences $\{a_j\}$ and $\{b_j\}$ as $\gamma_j = (a_j, b_j)$. Though the definition of the Korobov space $H_{d,\alpha,\gamma}$ uses only the truncated sequences $\{a_j\}_{j=1}^d, \{b_j\}_{j=1}^d$, our main results stated in Theorems 2 and 3 below give the dimension-independent rates of approximation and allow d to be as large as we want.

For $h = [h_1, h_2, \dots, h_d] \in \mathbb{Z}^d$, define the weight function as

$$\omega_{\alpha}(\gamma, h) = \prod_{j=1}^{d} \omega_{\alpha}(\gamma_j, h_j).$$

For the polynomial case with $\alpha < \infty$, let $b_j = 1$ for all j and

$$\omega_{\alpha}(\gamma_j, h_j) = \begin{cases} 1 & \text{if } h_j = 0 \\ |h_j|^{\alpha}/a_j & \text{if } h_j \ne 0. \end{cases}$$

For $\alpha = \infty$ in the exponential case, we fix $\omega > 1$, and take

$$\omega_{\infty}(\gamma_j, h_j) = \omega^{|h_j|^{b_j}/a_j}.$$

The Korobov space $H_{d,\alpha,\gamma}$ of complex-valued one-periodic functions is defined on $\mathbb{T}^d$ with a reproducing kernel

$$K_{\alpha,\gamma}(x,y) = \sum_{h\in\mathbb{Z}^d} \frac{\exp(2\pi i h\cdot(x-y))}{\omega_{\alpha}(\gamma,h)}.$$

Note that in both the polynomial case and exponential case, the kernel is bounded, that is, $\kappa := \sup_{x\in\mathbb{T}^d} \sqrt{K_{\alpha,\gamma}(x,x)} < \infty$.

The inner product on the RKHS can be easily seen by the periodicity of functions as

$$\langle f, g\rangle_H = \sum_{h\in\mathbb{Z}^d} \omega_{\alpha}(\gamma,h)\hat{f}(h)\overline{\hat{g}(h)}, \ f, g \in H_{d,\alpha,\boldsymbol{\gamma}}$$

where $\hat{f}$ is the Fourier series of f, given by

$$\hat{f}(h) = \int_{\mathbb{T}^d} f(x)\exp(-2\pi i\ h\cdot x)\,dx$$

and $\overline{\hat{g}(h)}$ is the complex conjugate of the Fourier series $\hat{g}$.

The norm in the RKHS $H_{d,\alpha,\boldsymbol{\gamma}}$ is then given by

$$\|f\|_H = \left(\sum_{h\in\mathbb{Z}^d} \omega_{\alpha}(\gamma,h)\left|\hat{f}(h)\right|^2\right)^{1/2}.$$

Since $\omega_{\alpha}(\gamma,h) \ge 1$, we have

$$\|f\|_{L_2(\mathbb{T}^d)} \le \|f\|_H$$

for all $f \in H_{d,\alpha,\boldsymbol{\gamma}}$. Thus, $H_{d,\alpha,\boldsymbol{\gamma}} \subset L_2(\mathbb{T}^d)$ where $L_2(\mathbb{T}^d)$ denotes the space of square integrable one-periodic functions with norm $\|f\|_{L_2(\mathbb{T}^d)} = (\int_{\mathbb{T}^d} |f(x)|^2 dx)^{1/2}$.

Remark 1. Before exhibiting the approximation in the Korobov space $H_{d,\alpha,\boldsymbol{\gamma}}$, we would address the essential assumption $\alpha > 1$ to control the smoothness. For example, taking $d = 1$, if α is an even integer, then for any $f \in H_{d,\alpha,\boldsymbol{\gamma}}$, f is $\frac{\alpha}{2}$ times differentiable, and its kth derivative is absolutely continuous for $k = 1, \dots, \frac{\alpha}{2} - 1$ while the $\frac{\alpha}{2}$-th derivative belongs to $L_2(\mathbb{T}^d)$. In the exponential case with $\alpha = \infty$, the Korobov space $H_{d,\alpha,\boldsymbol{\gamma}}$ consists of periodic functions that are analytic. For more details, refer to Novak and Wozniakowski (2008). ■

As in learning theory (Smale & Zhou, 2007), the Korobov space can be understood by an integral operator approach which will enable us to derive projections onto finite-dimensional subspaces for discretization in operator learning. Define the inclusion mapping $\mathrm{id} : H_{d,\alpha,\boldsymbol{\gamma}} \to L_2(\mathbb{T}^d)$ given by $\mathrm{id}(f) = f$. Then $\|\mathrm{id}\| = 1$ because $\|f\|_{L_2(\mathbb{T}^d)} \le \|f\|_H$ and the equality holds for constant functions. Moreover, the adjoint operator of id is the integral operator $\mathrm{id}^* : L_2(\mathbb{T}^d) \to H_{d,\alpha,\boldsymbol{\gamma}}$ given by

$$\mathrm{id}^*(g) = \int_{\mathbb{T}^d} K_{\alpha,\gamma}(\cdot, x)\, g(x) dx.$$

Consider the positive and self-adjoint operator $T := \mathrm{id}^* \mathrm{id} : H_{d,\alpha,\boldsymbol{\gamma}} \to H_{d,\alpha,\boldsymbol{\gamma}}$. We have

$$\langle Tf, g\rangle_H = \langle \mathrm{id}(f), \mathrm{id}(g)\rangle_{L_2(\mathbb{T}^d)} = \langle f, g\rangle_{L_2(\mathbb{T}^d)}$$

which follows by the reproducing property that for any $f \in H_{d,\alpha,\boldsymbol{\gamma}}$,

$$\begin{aligned} Tf(x) &= \langle Tf, K_{\alpha,\gamma}(x,\cdot)\rangle_H = \langle f, K_{\alpha,\gamma}(x,\cdot)\rangle_{L_2(\mathbb{T}^d)} \\ &= \sum_{h\in\mathbb{Z}^d} \omega_\alpha^{-1}(\gamma,h)\hat{f}(h)\exp(2\pi i h\cdot x). \end{aligned}$$

Here $\{\omega_\alpha^{-1/2}(\gamma,h)\exp(2\pi i h\cdot)\}_{h\in\mathbb{Z}^d}$ is a set of eigenvectors of T and an orthonormal basis of $H_{d,\alpha,\boldsymbol{\gamma}}$. Since

$$\sum_{h\in\mathbb{Z}^d} \omega_\alpha^{-1}(\gamma,h) = \kappa^2 < \infty,$$

T is a trace class operator, which follows that id is a Hilbert–Schmidt operator and thus compact. Now we can state projections on the Korobov space for discretization in operator learning. This is done by keeping components with truncating the eigenvalues $\{\omega_\alpha^{-1}(\gamma,h)\}$.

For $\epsilon \in (0,1)$, let

$$R(\epsilon,d) := \{h \in \mathbb{Z}^d : \omega_\alpha^{-1}(\gamma,h) > \epsilon^2\}$$

be the set of selected eigenvalues and

$$A_{n,d}(f)(x) := \sum_{h\in R(\epsilon,d)} \hat{f}(h)\exp(2\pi i h\cdot x)$$

with $n = |R(\epsilon,d)|$. According to a general procedure about such projections described in Lemma 4 in the Appendix, we see that $A_{n,d}$ achieves an L_2 approximation error ϵ, that is, $\|A_{n,d}(f) - f\|_{L_2(\mathbb{T}^d)} \le \epsilon\|f\|_H$ for $f \in H_{d,\alpha,\boldsymbol{\gamma}}$.

2.2. Fourier neural operator

One of the main findings of this paper is the realization of the projection $A_{n,d}$ by deep FNOs induced by the activation function $\sigma : \mathbb{C} \to \mathbb{C}$ coupled by the rectified linear unit (ReLU) $\max(0,x)$ and defined as $\sigma(z) = \max(0,x) + i\max(0,y)$ for all $z = x + iy \in \mathbb{C}$. A core ingredient of an FNO layer is a finite impulse response (FIR) filter or kernel P. Here, we take the FIR range to be the frequency domain

$$[h]_m = \{h \in \mathbb{Z}^d : |h|_\infty \le m\} \text{ with } m \in \mathbb{N}.$$

Then we define the truncated Fourier coefficients $\mathcal{F}_m : L_2(\mathbb{T}^d) \to \mathbb{C}^{[h]_m}$ and inverse transform $\mathcal{F}_m^{-1} : \mathbb{C}^{(2m+1)^d} \to L_2(\mathbb{T}^d)$ respectively as

$$\mathcal{F}_m(v)(h) = \int_{\mathbb{T}^d} v(x)\exp(-2\pi i\ h\cdot x)\, dx, \text{ for } h \in [h]_m$$

$$\mathcal{F}_m^{-1}(\hat{v})(x) = \sum_{h\in[h]_m} \hat{v}_h \exp(2\pi i\ h\cdot x), \text{ for } \hat{v} \in \mathbb{C}^{(2m+1)^d}.$$

An FIR kernel $P : [h]_m \to \mathbb{C}$ induces a filter operation on the space of periodic functions as $v \in L_2(\mathbb{T}^d) \to \mathcal{F}_m^{-1}\big(P \odot \mathcal{F}_m(v)\big)$ where $\odot$ is the Hadamard product given by

$$P \odot \mathcal{F}_m(v) = [P(h)\mathcal{F}_m(v)(h)]_{h\in[h]_m}.$$

Definition 1 (*Fourier Neural Operator*)**.** A deep FNO network $\{v_l\}_{l=0}^L$ of depth $L \in \mathbb{N}$ with $v_0 \in L_2(\mathbb{T}^d)$ is defined as

$$v_{l+1} = \sigma\Big(w_{l+1}v_l + c_{l+1} + \mathcal{F}_m^{-1}\big(P_{l+1} \odot \mathcal{F}_m(v_l)\big)\Big)$$

where $w_{l+1}, c_{l+1} \in \mathbb{C}$ are parameters and $P_{l+1} : [h]_m \to \mathbb{C}$ is a FIR kernel. For the filter sequence $\{P_l\}_{l=1}^L$, we define the kernel size $s \in \mathbb{N}$ to be $\max_{l\in L} |\operatorname{supp}(P_l)|$ where $\operatorname{supp}(P_l) = \{h \in [h]_m : P_l(h) \neq 0\}$ and $|\operatorname{supp}(P_l)|$ denotes its cardinality.

From Lemma 4 in the appendix, we know that the linear approximation $A_{n,d}$ is the optimal approximation to achieve an error bound ϵ with $n = |R(\epsilon,d)|$. We denote the radius of $R(\epsilon,d)$, to be the largest coordinate in norm, by

$$\tilde{R}(\epsilon,d) = \max_{h\in R(\epsilon,d)} |h_j|.$$

We define our L_2-approximation in the Korobov space $H_{d,\alpha,\boldsymbol{\gamma}}$ by the FNO layers as a mapping $\Psi : H_{d,\alpha,\boldsymbol{\gamma}} \to H_{d,\alpha,\boldsymbol{\gamma}}$ of the form

$$\Psi(f) := \mathcal{L}_L \circ \mathcal{L}_{L-1} \circ \cdots \circ \mathcal{L}_1(f)$$

where for $1 \le l \le L-1$, $\mathcal{L}_l(v_{l-1}) = v_l$ as defined in the iteration in Definition 1 and $v_0 = f$. For $l = L$, we add an end-to-end skip connection into the structure and remove the activation function:

$$\mathcal{L}_L(v_{L-1}) = w'_L v_0 + w''_L v_{L-1} + c_L + \mathcal{F}_m^{-1}\big(P_L \odot \mathcal{F}_m(v_{L-1})\big).$$

Then we have the following proposition proved in the appendix to show the capacity of FNOs to extract features for the approximation.

Proposition 1. *For the Korobov space $H_{d,\alpha,\boldsymbol{\gamma}}$ with $\alpha > 1$ and any $\epsilon \in (0,1)$, there exists a deep FNO Ψ with the kernel size $1 \le s \le |R(\epsilon,d)|$, $\lceil \frac{|R(\epsilon,d)|}{s} \rceil$ layers and $m = \tilde{R}(\epsilon,d)$, such that for any $f \in B_{d,\alpha,\boldsymbol{\gamma}}$, $\Psi(f) = A_{n,d}(f)$ and then $\|f - \Psi(f)\|_{L_2(\mathbb{T}^d)} \le \epsilon$, where $B_{d,\alpha,\boldsymbol{\gamma}} := \{f \in H_{d,\alpha,\boldsymbol{\gamma}} : \|f\|_H \le 1\}$.*

The FNO layers project the Korobov function space onto a finite-dimensional space with an estimation on L_2 error, instead of utilizing a fixed polynomial system on $L^2([-1,1]^d)$ in Song, Fan et al. (2023), Song, Liu et al. (2023). Korobov spaces exhibit different underlying structures characterized by the weight functions $\omega_\alpha(\gamma,h)$, which necessitates the use of varying systems of n trigonometric polynomials for the nth optimal approximations (Lemma 4). There is also an advantage of FNO layers over the fixed trigonometric systems to extract finite-dimensional features: the FNO layers with the same network structure can adapt to Korobov spaces with different structures and achieve the optimal approximations as described in Proposition 1.

For Ψ satisfying Proposition 1, the features extracted by the FNO Layers are defined as

$$\mathbf{F}_n(f) = W_{\mathbf{F}}\mathcal{F}'_m(\Psi(f)) \tag{2}$$

where $W_{\mathbf{F}}$ is a $2n \times 2(2m+1)^d$ matrix with $2n$ real weights and $\mathcal{F}'_m := \mathrm{vec} \circ \mathcal{F}_m$ with

$$\mathrm{vec}([x_j + iy_j]_{j=1}^t) = [x_1, y_1, \ldots, x_j, y_j, \ldots, x_t, y_t]^T \text{ for any } t \in \mathbb{N}.$$

The effect of the matrix $W_{\mathbf{F}}$ is to preserve information in certain frequency domain i.e., $h \in R(\epsilon,d)$ and filter out the other by taking

$$W_{\mathbf{F}}[\mathrm{Re}(\hat{f}(h)), \mathrm{Im}(\hat{f}(h))]_{h\in[h]_m}^T := [\mathrm{Re}(\hat{f}(h)), \mathrm{Im}(\hat{f}(h))]_{h\in R(\epsilon,d)}^T \in \mathbb{R}^{2n}.$$

Remark 2. Compared to recent advances in complex-valued neural networks (Geuchen & Voigtlaender, 2024; Voigtlaender, 2023), the FNO layers $\mathcal{F}_m \circ \Psi$ also produce complex vectors. However, the input of our FNO layers consists of functions from Korobov spaces rather than complex vectors in $\mathbb{C}^d$. In this setting, the analytic properties of complex functions become less significant when studying the approximation of Lipschitz functionals. Although exploring the use of an FNO structure for approximating complex functions would be interesting, it lies beyond the scope of this paper.

2.3. Fourier functional network

Based on the extracted Fourier features, we now utilize a DCNN with multiple channels to learn an approximation to nonlinear operators. The convolution operation (1) and downsampling methods were extensively studied for a DCNN network structure in Zhou (2020a, 2020b). Inspired by these works on DCNNs, we define multi-channel CNNs with downsampling operations here.

Definition 2 (*Convolution with Multiple Channels*)**.** For channel size $c \in \mathbb{N}$, let $\omega := \{w^{(j)}\}_{j=1}^{c}$ be a collection of convolutional kernels supported in $\{0, 1, \dots, s_c\}$. Then for an input sequence $x = (x_k)_{k\in\mathbb{Z}}$ with $x_k \in \mathbb{R}$ supported in $\{1, 2, \dots, d\}$, the 1-D multi-channel convolution with a replication padding $[\cdot]$ between ω and x is a summation of the convolutions on each channel, defined as

$$\omega * [x] = \sum_{j=1}^{c} \sigma\left(w^{(j)} * [x] + b_j\right)$$

where $b_j \in \mathbb{R}^{d+s_c+2}$,

$$[x] = (x_1, x_1, x_2, x_3, \dots, x_{d-1}, x_d, x_d)$$

and $w^{(j)} * [x]$ is defined to be the vector restricting onto $\{1, \dots, d+s\}$ of the sequence filter by (1) for $1 \le j \le c$.

Definition 3 (*Downsampled DCNN with Multiple Channels*)**.** For $D \in \mathbb{N}$, the downsampling operator $\mathcal{D}_v : \mathbb{R}^D \to \mathbb{R}^{\mathbf{k}(D)}$ with a scaling parameter $v \le D$ is defined as $\mathcal{D}_v(v) = (v_{kv+1})_{1\le k\le \mathbf{k}(D)}, v \in \mathbb{R}^D$, with

$$\mathbf{k}(D) := \max\left\{k : k \le \left\lfloor \frac{D}{v} \right\rfloor, kv + 1 < D\right\}.$$

Then a downsampled DCNN with multichannel kernels $\{\omega_l\}_{l=1}^{L}$ with kernel size s_c has widths $\{d_l\}_{l=0}^{L}$ defined iteratively by $d_0 = d'$ and $d_l = \mathbf{k}(d_{l-1} + s_c + 2)$ for $l = 1, \dots, L$, and is a sequence of function vectors

$$v_l(x) = \mathcal{D}_v(\omega_l * [v_{l-1}(x)] + b_l) \text{ for } 1 \le l \le L$$

where $b_l \in \mathbb{R}^{d_{l-1}+s_c+2}$ and $v_0(x) = x$.

Motivated by an interpolation framework in Song, Fan et al. (2023), the CNN structure defined above can realize a high-dimensional interpolation function with a great reduction in the number of parameters as shown in Lemma 1 below, which plays a significant role in the functional approximation.

Lemma 1. *Let $x = (x_1, x_2, \dots, x_d) \in [-1, 1]^d$, then the function $\min(x) = \min\{x_1, \dots, x_d\}$ can be represented by a DCNN of kernel size 2 and channel size 4 with $\lceil \log_2 d \rceil$ layers and $13\lceil \log_2 d \rceil$ total parameters.*

Then we shall present the definition of our novel Fourier Functional Network. The Fourier Functional Network consists of FNO layers followed by a multichannel DCNN, where FNO layers learn a vector representation for each input function from a Korobov space and the multichannel CNN performs an approximation to functions from a Euclidean space to $\mathbb{R}$.

Definition 4. Let $f \in L_2(\mathbb{T}^d)$. First we define for each $j \in \mathbb{N}$ that

$$\Phi^{(j)}(f) = \Gamma_L \sigma(W_j \mathbf{F}_n(f) + b_j)$$

where $\mathbf{F}_n$ is defined by (2), W_j is a $(4n^2+2n)\times 2n$ matrix with $8n^2$ possibly nonzero weights, $b_j \in \mathbb{R}^{4n^2+2n}$, $\Gamma_L : \mathbb{R}^{4n^2+2n} \to \mathbb{R}$ is a DCNN of kernel size 2 and channel size 4 with scaling parameter $v = 2$ in Lemma 1. Then for $\mathcal{M} \in \mathbb{N}$, the Fourier Functional Network Φ is defined as

$$\Phi(f) = \sum_{j=1}^{\mathcal{M}} \zeta_j \Phi^{(j)}(f) \tag{3}$$

where $\zeta_j \in \mathbb{R}$ for $1 \le j \le \mathcal{M}$.

3. Main results

Our main results are stated below and proved in the next section. Suppose $F : L_2(\mathbb{T}^d) \to \mathbb{R}$ is a continuous functional with modulus of continuity ω_F defined for $t > 0$ by

$$\omega_F(t) = \sup\{|F(f) - F(g)| : \|f - g\|_{L_2(\mathbb{T}^d)} \le t\}.$$

Then for the unit ball $B_{d,\alpha,\boldsymbol{\gamma}}$, we consider the functional approximation on the subspace $H_{d,\alpha,\boldsymbol{\gamma}}$ of $L_2(\mathbb{T}^d)$ by a Fourier Functional Network Φ. The following theorem demonstrates a general result of the approximation to a nonlinear continuous functional F.

Theorem 1. *Let $d, n, M \in \mathbb{N}$ and $n \ge 2$. If $F : L_2(\mathbb{T}^d) \to \mathbb{R}$ is a continuous functional with the modulus of continuity ω_F, then for the unit ball $B_{d,\alpha,\boldsymbol{\gamma}}$ of $H_{d,\alpha,\boldsymbol{\gamma}} \subset L_2(\mathbb{T}^d)$, there exists a Fourier Functional Network Φ with the number of possibly nonzero parameters at most $2M$ such that*

$$\sup_{f\in B_{d,\alpha,\boldsymbol{\gamma}}} |F(f) - \Phi(f)| \le \omega_F\left(\epsilon_{n,d}\right) + 4n\ \omega_F\left[C\left(\frac{n}{\sqrt{M}}\right)^{\frac{1}{n}}\right]$$

where $\epsilon_{n,d} := \sup_{f\in B_{d,\alpha,\boldsymbol{\gamma}}} \|f - A_{n,d}(f)\|_{L_2(\mathbb{T}^d)}$ and C is a constant independent of n or M.

To derive asymptotics of the approximation, we need to investigate properties of the weight function $\omega_\alpha(\gamma, h)$ which yield the approximation bound for $\epsilon_{n,d}$. The polynomial and exponential cases are dealt with in the following lemmas from Dick, Kritzer, Pillichshammer, and Wo'zniakowski (2014), Novak, Sloan, and Wo'zniakowski (2004).

Lemma 2. ***Polynomial Case:*** *If $1 < \alpha < \infty$, we define the sum-exponent s_a of the sequence $\{a_j\}$ as*

$$s_a = \inf\left\{s > 0 : \sum_{j=1}^{\infty} a_j^s < \infty\right\}.$$

If $s_a < \infty$, then for any positive η, there exists a positive A_η depending on η, α and $\{a_j\}$ such that

$$|R(\epsilon, d)| \le A_\eta \epsilon^{-(p^*+\eta)} \text{ for all } \epsilon \in (0, 1), \tag{4}$$

where $p^* = 2\max(s_a, \alpha^{-1})$ is so-called the exponent of strong tractability.

Lemma 3. ***Exponential Case:*** *If $\alpha = \infty$, then for any arbitrary sequence $\{a_j\}$ and $\{b_j\}$ with $B(d) := \sum_{j=1}^{d} \frac{1}{b_j}$ and $p_d^* := 1/B(d)$, we have the following exponential convergence:*

$$\sup_{f\in B_{d,\alpha,\boldsymbol{\gamma}}} \left\|f - A_{n,d}(f)\right\|_{L_2(\mathbb{T}^d)} \le A_0 \exp\left\{-\left(\frac{n}{A_1}\right)^{p_d^*}\right\} \tag{5}$$

where A_0, A_1 are constants, depending on d.

If $B := \sum_{j=1}^{\infty} \frac{1}{b_j} < \infty$ and $p^* = 1/B$, then the uniform exponential convergence can be obtained by (5) that

$$\sup_{f\in B_{d,\alpha,\boldsymbol{\gamma}}} \left\|f - A_{n,d}(f)\right\|_{L_2(\mathbb{T}^d)} \le A_0 \exp\left\{-\left(\frac{n}{A_1}\right)^{p^*}\right\}. \tag{6}$$

Then by Lemmas 2 and 3, the approximation rates, which beat the curse of dimension, are obtained from the general result in Theorem 1, and stated in the following:

Theorem 2. ***(Polynomial Case)*** *Let $d \in \mathbb{N}, M \ge 8, \beta > 0$ and $1 < \alpha < \infty$. Suppose the parameters of the weight function $\omega_\alpha(\gamma, h)$ satisfy that $s_a < \infty$. Let $p^* = 2\max(s_a, \alpha^{-1})$. If $\omega_F(r) \le \beta r^\lambda$ for some $\lambda \in (0, 1]$, then there exists a Fourier Functional Network with the number of possibly nonzero parameters at most $2M$ such that for any positive number η, with $p = p^* + \eta$, we have*

$$\sup_{f\in B_{d,\alpha,\boldsymbol{\gamma}}} |F(f) - \Phi(f)| = O\left(\left(\frac{\log M}{\log(\log M)}\right)^{-\lambda/p}\right).$$

Theorem 3. ***(Exponential Case)*** *Let $d, M \in \mathbb{N}, \beta > 0$ and $\alpha = \infty$. Suppose the parameters of the weight function $\omega_\alpha(\gamma, h)$ satisfy that $B = \sum_{j=1}^{\infty} \frac{1}{b_j} < \infty$. Let $p^* = 1/B$. If $\omega_F(r) \le \beta\, r^\lambda$ for some $\lambda \in (0, 1]$ and $M \ge N'_{\lambda,d,p^*} \in \mathbb{N}$ (to be given in the proof), then there exists a Fourier Functional Network with the number of possibly nonzero parameters $2M$ such that*

$$\sup_{f \in B_{d,\alpha,\gamma}} |F(f) - \Phi(f)| = O\left(\exp\left(-\tau(\log M)^{\frac{p^*}{p^*+1}}\right)\right)$$

where τ is a constant depending on d, λ, p^*.

4. Proof of main results

4.1. Error decomposition

The approximation of the nonlinear functional F will be analyzed by an error decomposition procedure. For any $f \in B_{d,\alpha,\gamma}$, we have the following error decomposition

$$\begin{aligned} |F(f) - \Phi(f)| &\le |F(f) - F \circ A_{n,d}(f)| + |F \circ A_{n,d}(f) - \Phi(f)| \\ &\le \omega_F\left(\left\|f - A_{n,d}(f)\right\|_{L_2(\mathbb{T}^d)}\right) + \left|\phi_F \circ (\mu_n \circ A_{n,d})(f) - \Phi(f)\right| \end{aligned} \tag{7}$$

with $\phi_F = F \circ \mu_n^{-1}$ where

$$\mu_n : (A_{n,d}(B_{d,\alpha,\gamma}), \|\cdot\|_{L_2(\mathbb{T}^d)}) \to (\mathbb{R}^{2n}, \|\cdot\|_2)$$

is an isometric isomorphism given by

$$\mu_n\left(\sum_{h \in R(\epsilon,d)} \hat{f}(h) \exp(2\pi i h \cdot)\right) = [\mathrm{Im}(\hat{f}(h)), \mathrm{Re}(\hat{f}(h))]_{h \in R(\epsilon,d)}.$$

For the RHS of (7), the first term can be bounded by the approximation error of the projection operator $A_{n,d}$. For the second term, note that $\phi_F : \mathbb{R}^{2n} \to \mathbb{R}$ is a real-valued function and we can estimate the modulus of continuity ϕ_F by ω_ϕ. In fact, for any $u, v \in \mathbb{R}^{2n}$, we have

$$\begin{aligned} |\phi_F(u) - \phi_F(v)| &= \left|F \circ \mu_n^{-1}(u) - F \circ \mu_n^{-1}(v)\right| \\ &\le \omega_F\left(\left\|\mu_n^{-1}(u) - \mu_n^{-1}(v)\right\|_{L_2(\mathbb{T}^d)}\right). \end{aligned}$$

Since $\|g\|_{L_2(\mathbb{T}^d)} = \left(\sum_{h \in \mathbb{Z}^d} |\hat{g}(h)|^2\right)^{1/2}$ for $g \in L_2(\mathbb{T}^d)$, the RHS of the above inequality becomes

$$\omega_F\left(\left(\sum_{k=1}^{n} (u_k^{\mathrm{Im}} - v_k^{\mathrm{Im}})^2 + (u_k^{\mathrm{Re}} - v_k^{\mathrm{Re}})^2\right)^{1/2}\right) = \omega_F(\|u - v\|_2)$$

which follows that $\omega_\phi(r) \le \omega_F(r)$.

Having obtained the smoothness condition of ϕ_F, we construct a multichannel DCNN by Lemma 1 to approximate ϕ_F in Theorem 1.

4.2. Proof of Lemma 1

Proof. We adopt a Divide and Conquer method here by noticing that $\min(x) = \min\{g(x)\}$ where

$$g(x) = \begin{cases} (\min(x_1, x_2), \min(x_3, x_4), \dots, \min(x_{d-1}, x_d)), & \text{if } d \text{ is even} \\ (\min(x_1, x_2), \min(x_3, x_4), \dots, \min(x_{d-2}, x_{d-1}), x_d), & \text{if } d \text{ is odd.} \end{cases}$$

Note that $\min(x_i, x_j) = -\max(-x_i, -x_j) = -\left(\frac{-x_i - x_j}{2} + \frac{\left|-x_i + x_j\right|}{2}\right)$ and that $\left|-x_j + x_i\right| = \sigma(x_j - x_i) + \sigma(x_i - x_j)$. Then it can be easily obtained that $\max(-x_j, -x_i)$ can be represented as a convolution operation with size 2 kernels and 4 channels. Let $\omega = \{w^{(1)}, w^{(2)}, w^{(3)}, w^{(4)}\}$ with $\omega^{(1)} = \left(-\frac{1}{2}, 0\right), \omega^{(2)} = \left(0, -\frac{1}{2}\right), \omega^{(3)} = \left(-\frac{1}{2}, \frac{1}{2}\right)$ and $\omega^{(4)} = \left(\frac{1}{2}, -\frac{1}{2}\right)$ and bias vectors $b_1 = \mathbf{1}, b_2 = \mathbf{1}, b_3 = \mathbf{0}$ and $b_4 = \mathbf{0}$. Then

$$\sigma(\omega^{(1)} * [x] + b_1) = \left(1, 1 - \frac{x_1}{2}, 1 - \frac{x_1}{2}, 1 - \frac{x_2}{2}, 1 - \frac{x_3}{2}, \dots,\right)$$

$$\sigma(\omega^{(2)} * [x] + b_2) = \left(1 - \frac{x_1}{2}, 1 - \frac{x_1}{2}, 1 - \frac{x_2}{2}, 1 - \frac{x_3}{2}, \dots,\right)$$

$$\sigma(\omega^{(3)} * [x] + b_3) = \left(\sigma\left(\frac{x_1}{2}\right), 0, \sigma\left(\frac{x_2 - x_1}{2}\right), \sigma\left(\frac{x_3 - x_2}{2}\right), \dots,\right)$$

$$\sigma(\omega^{(4)} * [x] + b_4) = \left(\sigma\left(-\frac{x_1}{2}\right), 0, \sigma\left(\frac{x_1 - x_2}{2}\right), \sigma\left(\frac{x_2 - x_3}{2}\right), \dots,\right).$$

It is easy to notice that starting from the third element, the sums of the elements at odd positions make up the negatives of desired minimums. These values can be retrieved by the downsampling operator with scaling parameter $\nu = 2$. With the replication padding, the last minimum sampled is always $\min(x_d, x_d)$ or $\min(x_{d-1}, x_d)$, instead of the value $\min(x_d, 0)$ at the end of vectors. No matter whether d is even or odd, the minimums of every two neighbor elements defined above are the elements of the output vector of $-\mathcal{D}_2(\omega * x + b)$ with $b = -\mathbf{2}$. Next, we show that the above CNN with $\lceil \log_2(d) \rceil$ layers can represent the minimum function.

The input dimension can be always written as $d = 2^p + q$ where $p, q \in \mathbb{N}$ and $q < 2^p$. Recall that the expression of the downsampling operator is $\mathcal{D}_2(v) = (v_{2k+1})_{1 \le k \le \mathbf{k}}$ where $\mathbf{k} = \lfloor \frac{D+1}{2} \rfloor$ for $v \in \mathbb{R}^{D+3}$.

Given $p', q' \in \mathbb{N}$ such that $q' < 2^{p'}$. Assume that for any $p \le p', q \le q'$ with $q < 2^p$, the minimum of $2^p + q$ elements can be obtained by the above CNN with $\lceil \log_2(2^p + q) \rceil$, i.e., $p + 1$ layers.

For $p = p' + 1, q = q'$, we know that after one layer of the CNN with downsampling, the dimension is reduced to $\left\lfloor \frac{2^{p'+1} + q' + 1}{2} \right\rfloor = 2^{p'} + \lfloor \frac{q'+1}{2} \rfloor$. Then the minimum can be obtained with another p' layers due to $\lfloor \frac{q'+1}{2} \rfloor \le q' < 2^{p'}$. Thus for $D = 2^{p'+1} + q'$, we can get the minimum by the CNN with $p' + 1$ layers.

For $p = p', q = q' + 1$, if $q' + 1 = 2^{p'}$, it corresponds to the case where $p = p' + 1$ and $q = 0$, under which the minimum can be retrieved with the CNN of $p' + 1$ layers. Otherwise, after one layer of the CNN, the dimension is reduced to $2^{p'-1} + \lfloor \frac{q'}{2} + 1 \rfloor$. Then the minimum can be obtained with another $p' - 1$ layers, since $q' < 2^{p'} - 1$. Thus, the CNN of p' layers can get the minimum with the input dimension $2^{p'} + q' + 1$.

By induction, we show that for any input vector with dimension d, the minimum of elements can be obtained with the CNN of $\lceil \log_2(d) \rceil$ layers. ■

4.3. Proof of Theorem 1

Proof. Similarly with Proposition 2 in Song, Fan et al. (2023), to construct an approximation to ϕ_F by a multichannel CNN in our Fourier functional network, we define the piecewise linear interpolation $H : \mathbb{R}^{2n} \to \mathbb{R}$ as

$$H(y) = \sum_{\xi \in \mathcal{G}} \phi_F(\xi) \psi\left(\frac{N}{2}(y - \xi)\right) \tag{8}$$

where $\mathcal{G} = \left\{-1 + \frac{2}{N} i : i = 0, \dots, N\right\}^{2n}$ and

$$\psi(y) = \sigma\left(\min\left\{\min_{k \ne j}\left(1 + y_k - y_j\right), \min_k\left(1 + y_k\right), \min_k\left(1 - y_k\right)\right\}\right). \tag{9}$$

We first denote a simplex in $\mathbb{R}^t$ by

$$\triangle_{\boldsymbol{\eta},\rho} = \left\{\mathbf{y} \in \mathbb{R}^{2n} : 0 \le y_{\rho(1)} - n_{\rho(1)} \le \cdots \le y_{\rho(2n)} - n_{\rho(2n)} \le 1\right\},$$

where $\boldsymbol{\eta} = (\eta_1, \dots, \eta_{2n}) \in \mathbb{Z}^{2n}, \mathbf{y} = (y_1, \dots, y_{2n}), \rho \in \mathcal{P}_{2n}$, the set of all permutations of $2n$ elements. Then $\left\{\triangle_{\boldsymbol{\eta},\rho}\right\}_{\boldsymbol{\eta} \in \mathbb{Z}^{2n}, \rho \in \mathcal{P}_{2n}}$ is a partition of $\mathbb{R}^{2n}$. It is easy to check that $\psi(\mathbf{0}) = 1$, and $\psi(\mathbf{y}) = 0$ for $\mathbf{y} \in \mathbb{Z}^{2n} \backslash \{\mathbf{0}\}$ and ψ is linear in each simplex $\triangle_{\boldsymbol{\eta},\rho}$ for $\boldsymbol{\eta} \in \mathbb{Z}, \rho \in \mathcal{P}_{2n}$.

By Lemma 1, we construct the linear interpolation ψ by multichannel CNNs with much fewer parameters. Since ψ is in the form of

$$\sigma(\min\{a_i : i = 1, \dots, 4n^2 + 2n\}) = \min\{\sigma(a_i) : i = 1, \dots, 4n^2 + 2n\},$$

then the piecewise linear interpolation $\psi : \mathbb{R}^{2n} \to \mathbb{R}$ can be represented by a DCNN of kernel size 2 and channel size 4 with $\lceil \log_2(4n^2 + 2n) \rceil$ layers and $13\lceil \log_2(4n^2 + 2n) \rceil$ in the form of

$$\psi(y) = -\Gamma_L(\sigma(W_2 y + b_2)) \tag{10}$$

where $L = \log_2(4n^2+2n)$, W_2 is a $(4n^2+2n)\times(2n)$ matrix with $8n^2$ nonzero weights $b_2 = \mathbf{1}$ such that

$$W_2 y + b_2 = [1+y_1, \dots, 1+y_{2n}, 1-y_1, \dots, 1-y_{2n}, \\ 1+y_1-y_2, \dots, 1+y_1-y_{2n}, \dots, 1+y_{2n}-y_1, \dots, 1+y_{2n}-y_{2n-1}]$$

Let $\psi_\xi = \psi\left(\frac{N}{2}(y-\xi)\right)$ and then it follows that

$$\psi_\xi = -\Gamma_L(\sigma(W_{2,N} y + b_{2,\xi}))$$

where $W_{2,N} = \frac{N}{2} W_2$ and $b_{2,\xi} = \mathbf{1} - W_{2,N}\xi$. Then H can be written into

$$H(y) = \sum_{\xi\in\mathcal{G}} (-\phi_F(\xi))\Gamma_L(\sigma(W_{2,N} y + b_{2,\xi})) \tag{11}$$

which can viewed as a linear combination of the outputs of the multi-channel CNN with different input features. This structured network has the depth $\lceil \log_2(4n^2+2n)\rceil$ and number of nonzero weights

$$M \le 2(N+1)^{2n} + 8n^2 + 13\lceil \log_2(4n^2+2n)\rceil \le c_1 n^2 (N+1)^{2n} \tag{12}$$

for some absolute constant c_1.

Next, we will show how to approximate ϕ_F by H and estimate the approximation error. For any $y \in [-1,1]^{2n}$, there exists a simplex $\triangle$ containing y. Then we denote the restriction of H on $\triangle$ by $H_\triangle$. Then for $r > 0$,

$$\omega_{H_\triangle}(r) = \sup\{|H_\triangle(y_1) - H_\triangle(y_2)| : \|y_1 - y_2\|_2 \le r, y_1, y_2 \in \triangle\} \\ \le \sup_{y\in\triangle} \|\nabla H_\triangle(y)\|_2 \, r \le \sqrt{2n} \sup_{y\in\triangle} |\nabla H_\triangle(y)|_\infty \, r \tag{13}$$

where ∇ is the gradient operator. Since $H_\triangle$ is linear in $\triangle$ and interpolates ϕ_F on every node, we can obtain that

$$|\partial_j H_\triangle(y)| = \left|\frac{N}{2}(\phi_F(\alpha_j) - \phi_F(\beta_j))\right| \le \frac{N}{2}\omega_\phi\left(\frac{2}{N}\right)$$

for $1 \le j \le 2n$, where α_j, β_j are the vertices of $\triangle$ with the same coordinates except for the jth coordinate. Since H coincides with ϕ_F on every vertex of $\triangle$, it follows that

$$\begin{aligned}\sup_{y\in[-1,1]^{2n}} |\phi_F(y) - H(y)| &\le \omega_\phi\left(\frac{\sqrt{2n}}{N}\right) + \omega_{H_\triangle}\left(\frac{\sqrt{2n}}{N}\right) \\ &\le \omega_\phi\left(\frac{\sqrt{2n}}{2}\cdot\frac{2}{N}\right) + n\omega_\phi\left(\frac{2}{N}\right) \\ &\le \omega_\phi\left(\lfloor\sqrt{2n}\rfloor\frac{2}{N}\right) + n\omega_\phi\left(\frac{2}{N}\right) \\ &\le \sqrt{2n}\omega_\phi\left(\frac{2}{N}\right) + n\omega_\phi\left(\frac{2}{N}\right) \\ &\le 4n\omega_\phi\left(\frac{2}{N}\right) \le 4n\omega_F(2/N).\end{aligned}$$

Since $N \ge \frac{M^{1/2n}}{c_1 n^{1/n}}$ from (12), we have that

$$\sup_{y\in[-1,1]^{2n}} \|\phi_F(y) - H(y)\| \le 4n\omega_F\left(\frac{c_2 n^{\frac{1}{n}}}{M^{\frac{1}{2n}}}\right) \tag{14}$$

where $c_2 = 2c_1$.

Let our Fourier functional network $\Phi_{\mathcal{M}}$ with $\mathcal{M} = (N+1)^{2n}$ to be

$$H(\mathbf{F}_n(f)) = \sum_{\xi\in\mathcal{G}} (-\phi_F(\xi))\Gamma_L(\sigma(W_{2,N}\mathbf{F}_n(f) + b_{2,\xi})). \tag{15}$$

Then from (7) and Proposition 1, it can be obtained that

$$\sup_{f\in B_{d,\alpha,\gamma}} |F(f) - \Phi(f)| \le \omega_F(\epsilon_{n,d}) + 4n\omega_F\left(\frac{Cn^{\frac{1}{n}}}{M^{\frac{1}{2n}}}\right). \tag{16}$$

Take M to be the largest number satisfying (12), then the total number of parameters $4n+M$ can be bounded by $2M$. The proof of Theorem 1 is complete. ■

4.4. Proof of Theorem 2

Proof. With the assumption on the weight function $\omega_\alpha(\gamma, h)$ in the polynomial case of Lemma 1, we have $\epsilon_{n,d} \le C_\eta n^{-1/(p^*+\eta)}$ with $p^* = 2\max(s_a, \alpha^{-1})$ and then for any $p = p^* + \eta > p^*$, it follows that

$$\sup_{f\in B_{d,\alpha,\gamma}} |F(f) - \Phi(f)| \le \beta\left(C_\eta n^{-1/p}\right)^\lambda + 4\beta n\left(Cn^{1/n}M^{-1/(2n)}\right)^\lambda \\ \le \beta C_\eta^\lambda n^{-\lambda/p} + 4\beta C^\lambda n^{\lambda/n} n M^{-\lambda/(2n)}.$$

Since $n^{1/n} \le 3$ for all $n \in \mathbb{N}$, by taking $C_{\beta,\lambda,\eta} = 16\beta(3C^\lambda + C_\eta^\lambda)$ we have

$$\sup_{f\in B_{d,\alpha,\gamma}} |F(f) - \Phi(f)| \le C_{\beta,\lambda,\eta}\left(n^{-\lambda/p} + nM^{-\lambda/(2n)}\right).$$

To obtain a convergence rate with respect to the number of possibly nonzero parameters $2M$, we need to build some relations between n and M. We choose n to be the integer such that

$$C_{\lambda,p} n\log n \le \log M < C_{\lambda,p}(n+1)\log(n+1)$$

where $C_{\lambda,p} = 2(p^{-1} + \lambda^{-1}) > 2$. It follows that

$$-\frac{\lambda}{p}\log n \ge \log n - \frac{\lambda}{2n}\log M$$

and therefore, $n^{-\lambda/p} \ge nM^{-\lambda/(2n)}$. Then we have

$$\sup_{f\in B_{d,\alpha,\gamma}} |F(f) - \Phi(f)| \le 2C_{\beta,\lambda,\eta} n^{-\lambda/p}.$$

Since $n \ge 2$, then we have $n \le \frac{\log M}{C_{\lambda,p}\log n} \le \log M$, which follows that

$$\log M < C_{\lambda,p}(n+1)\log(n+1) \le 4C_{\lambda,p} n\log n \le 4C_{\lambda,p} n\log(\log M).$$

Finally, we can obtain the result that

$$\sup_{f\in B_{d,\alpha,\gamma}} |F(f) - \Phi(f)| \le 2C_{\beta,\lambda,\eta}(4C_{\lambda,p})^{\lambda/p}\left(\frac{\log M}{\log(\log(M))}\right)^{-\lambda/p}.$$

It is easy to see $4n \le M$ by the relation between n and M. Then the total number is less than $2M$. This completes the proof of Theorem 2. ■

4.5. Proof of Theorem 3

Proof. In the exponential case, $\epsilon_{n,d}$ has an exponential convergence rate as $C_0\exp\left(-\frac{n^{p^*}}{C_1}\right)$ where C_0, C_1 are constants depending on dimension d. Then we have

$$\sup_{f\in B_{d,\alpha,\gamma}} |F(f) - \Phi(f)| \le \beta\left(C_0\exp\left(-\frac{n^{p^*}}{C_1}\right)\right)^\lambda + 4\beta n\left(Cn^{1/n}M^{-1/(2n)}\right)^\lambda \\ \le \beta C_0^\lambda \exp\left(-\frac{\lambda n^{p^*}}{C_1}\right) + 4\beta C^\lambda n^{\lambda/n} n M^{-\lambda/(2n)}.$$

Let $C'_{\beta,\lambda,\eta} = 16\beta(3C^\lambda + C_0^\lambda)$ and then it is obtained that

$$\sup_{f\in B_{d,\alpha,\gamma}} |F(f) - \Phi(f)| \le C'_{\beta,\lambda,\eta}\left(\exp\left(-\frac{\lambda n^{p^*}}{C_1}\right) + nM^{-\lambda/(2n)}\right).$$

Similarly, we need to balance the two terms on the RHS. Here we choose n to be the smallest integer not less than $N_{p^*} \in \mathbb{N}$ such that

$$C_{\lambda,d} n^{p^*+1} \le \log M < C_{\lambda,d}(n+1)^{p^*+1}$$

where $C_{\lambda,d} = 2\left(\frac{1}{C_1} + \frac{1}{\lambda}\right) > 2$ and N_{p^*} is determined by p^* in the form of $N_{p^*} = \min\{n \in \mathbb{N} : n^{p^*} \ge \log n\}$. It follows that

$$\frac{2}{C_1} n^{p^*+1} + \frac{2}{\lambda} n\log n \le C_{\lambda,d} n^{p^*+1} \le \log M$$

and that $\exp\left(-\frac{\lambda n^{p^*}}{C_1}\right) \ge nM^{-\lambda/(2n)}$. Then we have

$$\sup_{f\in B_{d,\alpha,\gamma}} |F(f) - \Phi(f)| \le 2C'_{\beta,\lambda,\eta}\exp\left(-\frac{\lambda n^{p^*}}{C_1}\right).$$

Similarly, we also have $\log M < C_{\lambda,d}(n+1)^{p^*+1} \leq C_{\lambda,d}2^{p^*+1}n^{p^*+1}$, and then we can obtain the final result

$$\sup_{f\in B_{d,\alpha,\gamma}} |F(f)-\Phi(f)| \leq 2C'_{\beta,\lambda,\eta}\exp\left(-\frac{\lambda}{C_{\lambda,d,p^*}}(\log M)^{p^*/(p^*+1)}\right)$$

with $C_{\lambda,d,p^*} = C_1(C_{\lambda,d}2^{p^*+1})^{p^*/(p^*+1)}$, when M is bigger than N'_{λ,d,p^*} where $N'_{\lambda,d,p^*} = \min\{m \in \mathbb{N} : \log m \geq C_{\lambda,d}N_{p^*}^{p^*+1}\}$. This proves Theorem 3. ■

CRediT authorship contribution statement

Peilin Liu: Writing – original draft, Methodology. **Yuqing Liu:** Writing – original draft. **Xiang Zhou:** Writing – review & editing, Supervision. **Ding-Xuan Zhou:** Writing – review & editing, Supervision, Methodology.

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.

Acknowledgments

The work of Ding-Xuan Zhou is partially supported by the Australian Research Council under project DP240101919 and partially supported by InnoHK initiative, the Government of the HKSAR, China, and the Laboratory for AI-Powered Financial Technologies, Australia. Xiang ZHOU acknowledges the supported by General Research Funds from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. 11308121, 11318522, 11308323) and the NSFC/RGC Joint Research Scheme, Hong Kong [RGC Project No. N-CityU102/20 and NSFC Project No. 12061160462].

Appendix

A.1. Finite-dimensional projection

Lemma 4 (Novak & Wozniakowski, 2008). *$S : H \to G$ be a bounded linear operator between a Hilbert space H with $\dim(H) \geq n$ and another Hilbert space G. Let $\sigma_i = \sqrt{\lambda_i}$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ are the eigenvalues of $W = S^*S : H \to H$ with $W(e_i) = \lambda_i e_i$ and orthonormal $\{e_i\}$. Then the linear algorithm*

$$A_n(f) = \sum_{i=1}^{n} \langle f, e_i\rangle S(e_i)$$

is the nth optimal approximation.

A.2. Proof of Proposition 1

Proof. For the Korobov space $H_{d,\alpha,\gamma}$, given $\epsilon \in (0,1)$, consider the operator $A_{n,d}(f)(x) = \sum_{h\in R(\epsilon,d)} \hat{f}(h)\exp(2\pi i h\cdot x)$ where $R(\epsilon,d) = \{h \in \mathbb{Z}^d : \omega_\alpha^{-1}(\gamma,h) > \epsilon^2\}$. Recall $m = \tilde{R}(\epsilon,d)$. Then $R(\epsilon,d) \subset [h]_m$. Let $L = \lceil \frac{|R(\epsilon,d)|}{s} \rceil$. We can make a partition $\{\mathcal{P}_l\}_{l=1}^L$ of the set $R(\epsilon,d)$ such that

$$R(\epsilon,d) = \mathcal{P}_1 \cup \mathcal{P}_2 \cup \cdots \cup \mathcal{P}_L, \quad \mathcal{P}_j \cap \mathcal{P}_k = \emptyset \text{ if } j \neq k,$$

$$|\mathcal{P}_l| = s \text{ for } l \leq L-1, \quad \text{and } |\mathcal{P}_L| = |R(\epsilon,d)| - (L-1)s.$$

And we let $\mathbf{0} \in \mathcal{P}_1$. With the partition of $R(\epsilon,d)$, we have

$$A_{n,d}(f) = \sum_{l=1}^{L} A_{\mathcal{P}_l,d}(f)$$

where

$$A_{\mathcal{P}_l,d}(f)(x) = \sum_{h\in\mathcal{P}_l} \hat{f}(h)\exp(2\pi i h\cdot x).$$

Then we construct the Hadamard multiplication factors $\{P_l\}_{l=1}^L$ with the sets $\{\mathcal{P}_l\}_{l=1}^L$ of the partition by

$$P_l(h) = \mathbf{1}_{\mathcal{P}_l}(h) \text{ where } \mathbf{1}_{\mathcal{P}_l}(h) = \begin{cases} 1 & \text{if } h \in \mathcal{P}_l, \\ 0 & \text{if } h \in [h]_m \backslash \mathcal{P}_l. \end{cases}$$

For the other parameters in the FNO, we let $w_1 = \cdots = w_{L-1} = w''_L = 1$, $w'_L = -1$ and $c_1 = 2\kappa + 2\kappa i$, $c_l = 0$ for all $2 \leq l \leq L-1$ and $c_L = -2\kappa - 2\kappa i$. With the above parameterization, the part of Fourier (inverse) transform computation $\mathcal{F}_m^{-1}(P_l \odot \mathcal{F}_m(\cdot))$ in each layer is the orthogonal projection from $H_{d,\alpha,\gamma}$ to the subspace in the form of $\text{span}\{\exp(2\pi i\ h\cdot\ )\}_{h\in\mathcal{P}_l}$. We denote this part of each layer by $\text{Proj}_{\mathcal{P}_l}(\cdot)$. Then for $1 \leq l \leq L-1$, $v_0 = f$, then the iterative update becomes

$$v_l = \sigma\big(v_{l-1} + (2\kappa+2\kappa i)\delta_{1,l} + \text{Proj}_{\mathcal{P}_l}(v_{l-1})\big)$$

and the last layer is

$$v_L = -v_0 + v_{L-1} - 2\kappa - 2\kappa i + \text{Proj}_{\mathcal{P}_L}(v_{L-1}).$$

We claim that for $1 \leq l \leq L-1$, $v_l = f + (2\kappa+2\kappa i) + \sum_{k=1}^{l} A_{\mathcal{P}_k,d}(f)$.

For $l = 1$, it is easily obtained that $v_1 = \sigma\big(f + (2\kappa+2\kappa i) + A_{\mathcal{P}_1,d}(f)\big)$ and the activation function can be removed since

$$\Big|f(x) + A_{\mathcal{P}_1,d}(f)(x)\Big| \leq |f(x)| + |A_{\mathcal{P}_1,d}(f)(x)| \leq 2\kappa\,\|f\|_H \leq 2\kappa.$$

Therefore, the statement holds for $l = 1$.

Suppose the statement holds for $l \geq 1$. Then for the $(l+1)$-th layer, we have

$$v_{l+1} = \sigma\big(v_l + \text{Proj}_{\mathcal{P}_{l+1}}(v_l)\big), \tag{17}$$

$$v_l = f + (2\kappa+2\kappa i) + \sum_{k=1}^{l} A_{\mathcal{P}_k,d}(f). \tag{18}$$

Since $\mathbf{0} \in \mathcal{P}_1$ and $(\cup_{k=1}^{l}\mathcal{P}_k) \cap \mathcal{P}_{l+1} = \emptyset$, then $\text{Proj}_{\mathcal{P}_{l+1}}(v_l) = \text{Proj}_{\mathcal{P}_{l+1}}(f) = A_{\mathcal{P}_{l+1},d}(f)$. It follows that

$$v_{l+1} = \sigma\big(f + (2\kappa+2\kappa i) + \sum_{k=1}^{l} A_{\mathcal{P}_k,d}(f) + A_{\mathcal{P}_{l+1},d}(f)\big) \tag{19}$$

$$= f + (2\kappa+2\kappa i) + \sum_{k=1}^{l+1} A_{\mathcal{P}_k,d}(f). \tag{20}$$

Thus, the claim is proved.

For the last layer L, we have that

$$v_L = -f + v_{L-1} - 2\kappa - 2\kappa i + \text{Proj}_{\mathcal{P}_L}(v_{L-1}) \tag{21}$$

$$= \sum_{l=1}^{L} A_{\mathcal{P}_l,d}(f) = A_{n,d}(f). \tag{22}$$

This completes the proof of Proposition 1.

Data availability

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

References

Anandkumar, Anima, Azizzadenesheli, Kamyar, Bhattacharya, Kaushik, Kovachki, Nikola, Li, Zongyi, Liu, Burigede, et al. (2020). Neural operator: Graph kernel network for partial differential equations. In *ICLR 2020 workshop on integration of deep neural models and differential equations*.

Barron, Andrew R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, *39*(3), 930–945.

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

Dick, Josef, Kritzer, Peter, Pillichshammer, Friedrich, & Wo'zniakowski, Henryk (2014). Approximation of analytic functions in Korobov spaces. *Journal of Complexity*, *30*(2), 2–28.

Geuchen, Paul, & Voigtlaender, Felix (2024). Optimal approximation using complex-valued neural networks. *Advances in Neural Information Processing Systems*, *36*.

Karniadakis, George Em, Kevrekidis, Ioannis G., Lu, Lu, Perdikaris, Paris, Wang, Sifan, & Yang, Liu (2021). Physics-informed machine learning. *Nature Reviews Physics*, *3*(6), 422–440.

Klusowski, Jason M., & Barron, Andrew R. (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.

Kovachki, Nikola, Lanthaler, Samuel, & Mishra, Siddhartha (2021). On universal approximation and error bounds for Fourier neural operators. *Journal of Machine Learning Research*, *22*(290), 1–76.

LeCun, Yann, Bengio, Yoshua, & Hinton, Geoffrey (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Li, Zongyi, Kovachki, Nikola, Azizzadenesheli, Kamyar, Liu, Burigede, Bhattacharya, Kaushik, Stuart, Andrew, et al. (2020). Fourier neural operator for parametric partial differential equations. arXiv preprint arXiv:2010.08895.

Mao, Tong, & Zhou, Ding-Xuan (2022). Approximation of functions from Korobov spaces by deep convolutional neural networks. *Advances in Computational Mathematics*, *48*(6), 84.

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

Narhar Mhaskar, Hrushikesh (1993). Approximation properties of a multilayered feedforward artificial neural network. *Advances in Computational Mathematics*, *1*(1), 61–80.

Novak, Erich, Sloan, Ian H., & Wo'zniakowski, Henryk (2004). Tractability of approximation for weighted Korobov spaces on classical and quantum computers. *Foundations of Computational Mathematics*, *4*, 121–156.

Novak, E., & Wozniakowski, H. (2008). Tractability of multivariate problems. In *European math: vol. 2, Volume i: linear information*. (no. 3).

Petersen, Philipp, & Voigtlaender, Felix (2018). Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, *108*, 296–330.

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

Raissi, Maziar, Perdikaris, Paris, & Karniadakis, George E. (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.

Senior, Andrew W., Evans, Richard, Jumper, John, Kirkpatrick, James, Sifre, Laurent, Green, Tim, et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706–710.

Shi, Zhongjie, Yu, Zhan, & Zhou, Ding-Xuan (2023). Learning theory of distribution regression with neural networks. arXiv preprint arXiv:2307.03487.

Smale, Steve, & Zhou, Ding-Xuan (2007). Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, *26*(2), 153–172.

Song, Linhao, Fan, Jun, Chen, Di-Rong, & Zhou, Ding-Xuan (2023). Approximation of nonlinear functionals using deep ReLU networks. *Journal of Fourier Analysis and Applications*, *29*(4), 50.

Song, Linhao, Liu, Ying, Fan, Jun, & Zhou, Ding-Xuan (2023). Approximation of smooth functionals using deep ReLU networks. *Neural Networks*, *166*, 424–436.

Voigtlaender, Felix (2023). The universal approximation theorem for complex-valued neural networks. *Applied and Computational Harmonic Analysis*, *64*, 33–61.

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

Yu, Zhan, & Zhou, Ding-Xuan (2023). Deep learning theory of distribution regression with CNNs. *Advances in Computational Mathematics*, *49*(4), 51.

Zhou, Ding-Xuan (2020a). Theory of deep convolutional neural networks: Downsampling. *Neural Networks*, *124*, 319–327.

Zhou, Ding-Xuan (2020b). Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, *48*(2), 787–794.