Journal of Fourier Analysis and Applications (2026) 32:52
https://doi.org/10.1007/s00041-026-10260-4

Barron Space for Graph Convolutional Neural Networks

Seok-Young Chung[1] · **Qiyu Sun**[2]

Received: 26 February 2025 / Revised: 30 November 2025 / Accepted: 22 March 2026

Abstract
Graph convolutional neural network (GCNN) offers a framework for representing and learning functions of data on networks and irregular domains. In this paper, we introduce the concept of a graph Barron space of functions. We prove that the proposed graph Barron space is a reproducing kernel Banach space, can be decomposed into a union of reproducing kernel Hilbert spaces with explicitly expressed neuron kernels, and is dense in the space of continuous functions under certain technical assumptions. In this paper, we also show that the outputs of shallow GCNNs belong to the graph Barron space and that functions in the graph Barron space can be well approximated by outputs of shallow GCNN in both the integrated square and uniform norms. Moreover, we estimate the Rademacher complexity of functions with bounded Barron norm and conclude that functions in the graph Barron space can be learned efficiently from their noiseless random samples with high probability. Finally, we test the approximation performance of shallow GCNNs to a quadratic function on the data set collected at weather stations in the region of Brest, France.

Keywords Graph convolution neural network · Graph Barron space · Reproducing kernel Banach space · Reproducing kernel Hilbert space · Universal approximation · Radmacher complexity · Learnability

1 Introduction

Graph signal processing (GSP) provides an innovative approach to extract knowledge from massive datasets residing on networks and irregular domains [12, 16, 20, 28, 46,

Communicated by Isaac Pesenson.

✉ Qiyu Sun
qiyu.sun@ucf.edu

Seok-Young Chung
sychung@msu.edu

[1] Department of Mathematics, Michigan State University, East Lansing, MI 48824, USA

[2] School of Data, Mathematical, and Statistical Sciences, University of Central Florida, Orlando, FL 32765, USA

Published online: 06 May 2026

47, 52–54, 60, 64]. Many of these irregular structures could be modeled using graphs. For example, vertices of a graph can represent sensors in a sensor network with edges depicting peer-to-peer communication links, and similarly, the skeleton of a human body can be modeled as a graph, with joints as vertices and their natural anatomical connections as edges [1, 13, 59, 69, 77]. In this paper, we consider undirected graphs $\mathcal{G} := (V, E)$ of very large order $N \geq 1$ with the number of edges $\#E$ significantly less than $N(N-1)/2$. Ideally, the number of edges does not exceed a multiple of the graph order N, as is the case for graphs with bounded degree.

Convolutional neural network (CNN) has gained a lot of attention from industrial and academic communities, and it has made numerous achievements. For instance, computer vision based on CNNs makes it possible to accomplish tasks, such as face recognition, autonomous vehicle and intelligent medical treatment. The reader may refer to [26, 34, 35, 37, 40, 45, 74, 83, 84] and references therein for historical remarks and recent advances. Graph convolutional neural networks (GCNNs) generalize classical CNNs to process graph-structured data and have achieved strong performance across a wide range of tasks. However, a feasible extension of CNNs from regular grid (such as pixel grids to represent images) to irregular graph (such as the skeleton structure of human body) is not straightforward, and there is a significant gap between its theoretical foundations and engineering applications, see [6, 16, 28, 36, 39, 43, 60, 70, 81, 85] and references therein.

GCNN takes advantage of topological structure of the underlying graph and aggregates node information from the neighborhoods in a convolutional fashion. A basic question is how to define graph convolution appropriately. Two conventional approaches have been proposed to define graph convolution, one from the spectral perspective while the other from the spatial perspective. In the first trial to define a GCNN, the spatial convolution is used to sum up the neighboring features [8], see Section 7. In this paper, we adopt the spectral approach and use graph Fourier transform to define graph convolution $\mathbf{b} * \mathbf{x}$ between two graph signals $\mathbf{b}$ and $\mathbf{x}$, see (2.9).

Let $\mathbb{R}^V$ be the linear space of all graph signals, represented by column vectors $\mathbf{x} = (x(v))_{v\in V}$, on the graph $\mathcal{G} := (V, E)$, Ω be a compact subset of $\mathbb{R}^V$, $\mathcal{W} \subset \mathbb{R}^V$ be the linear space of graph signals $\mathbf{b}$ used for the convolution in GCNNs, $\sigma(t) = \max(0, t)$ be the ReLU activation function with $\sigma(\mathbf{y})$, $\mathbf{y} \in \mathbb{R}^V$, defined componentwisely, and denote the transpose of a vector $\mathbf{a} \in \mathbb{R}^V$ by $\mathbf{a}^T$. In this paper, we consider shallow (two-layer) GCNNs equipped with M neurons at every vertex $v \in V$, which has graph signal input $\mathbf{x}$ in the compact domain Ω and scalar-valued output given by

$$f_M(\mathbf{x}, \mathbf{\Theta}) = \frac{1}{M} \sum_{m=1}^{M} \mathbf{a}_m^T \sigma(\mathbf{b}_m * \mathbf{x} + \mathbf{c}_m), \ \mathbf{x} \in \Omega, \tag{1.1}$$

where $\mathbf{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$ and $\boldsymbol{\theta}_m = (\mathbf{a}_m, \mathbf{b}_m, \mathbf{c}_m) \in \mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V$, $1 \leq m \leq M$; see Section 2.3 for detailed description. The above shallow GCNN has neurons $\phi(\mathbf{x}, \boldsymbol{\theta}_m) = \mathbf{a}_m^T \sigma(\mathbf{b}_m * \mathbf{x} + \mathbf{c}_m)$, $1 \leq m \leq M$, and $2NM \leq (2N + \dim \mathcal{W})M \leq 3NM$ parameters, where $\dim \mathcal{W}$ is the dimension of the linear space $\mathcal{W}$. For the case that the convolution space $\mathcal{W}$ contains graph signals such that the corresponding convolution operation can be implemented by some polynomial filtering procedure with polynomials up to a given

degree L, i.e.,

$$\mathbf{b}_m * \mathbf{x} = h_m(\mathbf{S}_1, \ldots, \mathbf{S}_K)\mathbf{x}$$

for some commutative graph shifts $\mathbf{S}_1, \ldots, \mathbf{S}_K$ and multivariate polynomials h_m, $1 \le m \le M$, of degrees at most L, the above shallow GCNN is essentially the ChebNet in the literature [15, 28, 33, 43, 70], and it can be implemented in a distributed and **scalable** manner.

In this paper, we introduce the graph Barron space of functions on the domain Ω and address whether and how a function f in the Barron space can be approximated by the outputs f_M of shallow GCNNs, i.e.,

$$f(\mathbf{x}) \approx f_M(\mathbf{x}, \mathbf{\Theta}), \quad \mathbf{x} \in \Omega$$

for some appropriately chosen parameter $\mathbf{\Theta} \in (\mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V)^M$ and suitably large number M of neurons per vertex. Our **default** assumptions on the shallow GCNNs are that the underlying graph $\mathcal{G}$ has a very large order N, and the number of neurons M at each vertex should not exceed a multiple of the graph order N, and ideally, it is independent of the graph order so that the shallow GCNN remains scalable, see Section 4.1. In situations where shallow GCNNs require a large number of neurons at each vertex, it may be more advantageous to use deep GCNNs, where each layer contains a smaller number of neurons per vertex. Deep GCNNs remain relatively understudied, and we refer the reader to [40, 74, 83, 84] and the references therein for deep convolutional neural networks in the classical Euclidean setting.

We say that $A = O(B)$ for two quantities A and B if $|A/B|$ is bounded by some absolute constant. The main contributions of this paper are as follows.

(i) Barron space of functions on the unit cube $[0, 1]^N$ was introduced in [3] with the help of Fourier transform. In this paper, we follow the spatial framework in [2, 17–19, 63] and introduce a graph Barron space $\mathcal{B}$ on a compact domain Ω of (sparse) graph signals in $\mathbb{R}^V$. We show that the graph Barron space $\mathcal{B}$ is a reproducing kernel Banach space and functions in the Barron space $\mathcal{B}$ are Lipschitz continuous; see Theorem 3.1 and Corollary 3.4. We observe that functions in the graph Barron space $\mathcal{B}$, without the underlying graph structure into consideration, could be treated as a function on a domain of the Euclidean space $\mathbb{R}^N$, and belong to the Barron space defined in [17], however the converse may not be true, see Remarks 3.5 and 3.6.

(ii) Reproducing kernel Hilbert/Banach spaces (RKHSs/RKBSs) are ideal for function estimation, and their kernels are selected to measure certain similarity between input data [24, 51, 55, 65, 71]. For the graph Barron space $\mathcal{B}$, we do not have an explicit closed-form expression for its reproducing kernel. To address this, we decompose $\mathcal{B}$ into the union of a family of RKHSs equipped with neuron kernels, whose reproducing kernels have explicit expressions. We further establish norm equivalence between these spaces; see Theorems 3.10 and 3.12, and also Remark 3.13 for the comparison to the Barron space in the classical Euclidean setting.

(iii) The approximation of functions in Barron/Besov/Hölder spaces by outputs of some neural networks is well studied in the classical setting, see [17, 58, 61, 75,

76] and references therein. In this paper, we show that functions in the Barron space can be well approximated by some shallow GCNNs with bounded path norm, and conversely the limit of outputs of shallow GCNNs with bounded path norm belongs to the Barron space; see Theorems 4.1, 4.3 and 4.6. To achieve an integrated square (respectively, uniform) approximation error for a function f in the Barron space $\mathcal{B}$ using outputs of shallow GCNNs with accuracy $\epsilon\|f\|_{\mathcal{B}}$ on the domain $\Omega_{\rm sp}$ in (2.15), we derive from Theorems 4.1 and 4.3 that the number of parameters in the shallow GCNN is approximately $O(N\epsilon^{-2})$ (respectively, $O(Ns\epsilon^{-2}\log(s^{-1}\epsilon^{-1}N))$), and hence the shallow GCNN is nearly scalable. Here N is the order of the underlying graph $\mathcal{G}$ of the shallow GCNN, and $\|f\|_{\mathcal{B}}$ is the Barron norm of the function $f \in \mathcal{B}$. As expected, the approximation of functions in the Barron space does not suffer from the curse of dimensionality (i.e., the order N of the underlying graph $\mathcal{G}$ in the current setting), see Remarks 4.4 and 4.5 for comparisons in the classical neural network setting.

(iv) The universal approximation theorem provides a key theoretical justification for the design and use of neural networks [7, 31, 32, 49]. In Theorem 4.7, we establish the universal approximation theorem for shallow GCNNs and density of the graph Barron space $\mathcal{B}$, under some technical conditions on the graph Fourier transform and the convolution space $\mathcal{W}$.

(v) Rademacher complexity of a function class is a conventional measure of generalization error in learning theory [2, 4, 17, 57, 78]. In this paper, we provide an estimate to the Rademacher complexity of functions with bounded Barron norm, which depends on the inverse square root of the sample size and the square root of the logarithm of the graph order, see Theorem 5.1. As a consequence, we see that functions in the Barron space could be learnt from their random samples in an efficient way, see Theorem 5.4.

The rest of this paper is organized as follows. Graph shifts provide the foundation from which most GSP tools and techniques are derived. In Section 2.1, we recall some preliminaries on graph shifts $\mathbf{S}_1, \ldots, \mathbf{S}_K$, including their eigendecomposition in (2.3), polynomial filters in (2.7), and joint spectrum in Assumption 2.1. The graph Fourier transform (GFT) and graph convolution are key tools widely used in GSP. In Section 2.2, we recall the conventional definition of the GFT and the graph convolution based on the graph shifts, which is used in our GCNNs. In Section 2.3, we set up the essential components of our GCNNs, including the domain Ω, the activation function σ, the convolution norm $\|\cdot\|_{\rm co}$, and the formulation of shallow GCNNs. In Section 3.1, we introduce the graph Barron space $\mathcal{B}$ and show that it is a reproducing kernel Banach space. In Section 3.2, we introduce a family of reproducing kernel Hilbert spaces (RKHSs), provide explicit formulas for their reproducing kernels, and establish their integral representation. More importantly, in Section 3.3, we show that the graph Barron space $\mathcal{B}$ is the union of the RKHSs introduced in Section 3.2 with an associated norm equivalence. In Sections 4.1 and 4.2, we consider the approximation problem of functions in the Barron space by outputs of shallow GCNNs with bounded path norm in the integrated square norm and in the uniform norm respectively, and provide an estimate to number of parameters required for shallow GCNNs to reach a given

approximation accuracy. In Section 4.4, we establish a universal approximation theorem in the GCNN setting. In Section 5.1, we derive an upper bound for the Rademacher complexity of functions in the Barron space. In Section 5.2, we address the learnability of functions in the graph Barron space from their noiseless random sampling data. In Section 6, we use stochastic gradient descent with Nesterov momentum (SGDM) to train shallow GCNNs from both synthetic and real data, and we demonstrate its approximation performance in Sections 4 and 5. We observe that there is a trade-off between the number of neurons and the number of iterations in the SGDM that needs to be carefully balanced to achieve optimal performance of GCNNs. In the Conclusion and Discussions section, we consider a Barron space with convolution defined via the spatial approach (which involves far more parameters than the spectral approach in the paper) and discuss its approximation properties using the outputs of shallow GCNNs.

2 Preliminaries

In this paper, we consider undirected connected graphs $\mathcal{G} = (V, E)$ of very large order $N \geq 1$ with their adjacency, degree and Laplacian matrices denoted by $\mathbf{W}, \mathbf{D}$ and $\mathbf{L} := \mathbf{D} - \mathbf{W}$ respectively, and we define the *geodesic distance* $\rho(v, v')$ between vertices $v, v' \in V$ by the number of edges in the shortest path connecting them. For convenience, we use $\mathbf{x} = (x(v))_{v \in V}$ to denote a graph signal that takes value $x(v)$ at the vertex $v \in V$ and denote the set of all graph signals on the graph $\mathcal{G}$ by $\mathbb{R}^V$.

The concept of graph shifts is similar to the one-order delay in classical signal processing and polynomial filters have been widely used in graph signal processing. In Section 2.1, we recall some preliminaries on commutative graph shifts and polynomial filters [20, 23, 28, 29, 46, 47, 53, 54, 60, 64]. The graph Fourier transform is one of fundamental tools in graph signal processing that decomposes graph signals into different frequency components and represents them by different modes of variation [9, 11, 14, 28, 46, 47, 52, 60, 64]. Based on commutative graph shifts, in Section 2.2 we introduce graph Fourier transform of a graph signal and define graph convolution between two graph signals, see (2.5) and (2.9). We also show that the proposed graph convolution operation can be implemented in the spectral domain by taking the inverse Fourier transform of the multiplication between two Fourier transformed graph signals, and also in the spatial domain by applying some polynomial filtering procedure, see (2.9) and (2.10).

GCNNs are generalizations of classical CNNs to handle graph data, and they have been a powerful graph analysis method. In Section 2.3, we discuss the setting of a shallow GCNN on a compact domain of (sparse) graph signals.

2.1 Commutative Graph Shifts

A filter in graph signal processing is a linear transform to map one graph signal on the graph to another graph signal on the same graph, which is usually represented by a matrix with entries indexed by the vertex set. A *graph shift* $\mathbf{S}$, to be represented by

a matrix $\mathbf{S} = (S(v, v'))_{v,v' \in V}$, is an elementary graph filter satisfying

$$S(v, v') = 0 \text{ if } \rho(v, v') \geq 2. \tag{2.1}$$

Our illustrative examples of graph shifts are the degree matrix $\mathbf{D}$, the adjacency matrix $\mathbf{W}$, the Laplacian matrix $\mathbf{L}$, the symmetric normalized Laplacian matrix $\mathbf{L}^{\text{sym}} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$ and their variants [20, 23, 28, 46, 47, 53, 60, 64]. A significant advantage of a graph shift $\mathbf{S} = (S(v, v'))_{v,v' \in V}$ is that the filtering procedure $\mathbf{S} : (x(v))_{v \in V} =: \mathbf{x} \longmapsto \mathbf{S}\mathbf{x} := (\tilde{x}(v))_{v \in V}$ can be implemented by some local operation that updates signal value $\tilde{x}(v)$ at each vertex $v \in V$ by a "weighted" sum of signal values $x(v')$ at adjacent vertices $v' \in \mathcal{N}_v$,

$$\tilde{x}(v) = \sum_{v' \in \mathcal{N}_v} S(v, v')x(v'),$$

where $\mathcal{N}_v$ is the set of adjacent vertices of $v \in V$.

Similar to the one-order delay $z_1^{-1}, \ldots, z_K^{-1}$ in classical multidimensional signal processing, the concept of multiple commutative graph shifts $\mathbf{S}_1, \ldots, \mathbf{S}_K$ is introduced in [20], where two illustrative families of commutative graph shifts on circulant/Cayley graphs and product graphs are presented. Here we say that $\mathbf{S}_1, \ldots, \mathbf{S}_K$ are *commutative* if

$$\mathbf{S}_k\mathbf{S}_{k'} = \mathbf{S}_{k'}\mathbf{S}_k, \ 1 \leq k, k' \leq K. \tag{2.2}$$

For commutative graph shifts, it is well known that they can be upper-triangularized simultaneously by some unitary matrix [27, Theorem 2.3.3]. Under additional real-valued and symmetric assumptions, commutative graph shifts can be diagonalized simultaneously by some orthogonal matrix $\mathbf{U}$, i.e.,

$$\mathbf{S}_k = \mathbf{U}\mathbf{\Lambda}_k\mathbf{U}^T \tag{2.3}$$

for some diagonal matrices $\mathbf{\Lambda}_k := \text{diag}(\lambda_k(n))_{1 \leq n \leq N}$, $1 \leq k \leq K$. Define

$$\mathbf{\Lambda} = \left\{\boldsymbol{\lambda}(n) = [\lambda_1(n), ..., \lambda_K(n)]^T, 1 \leq n \leq N\right\} \subset \mathbb{R}^K. \tag{2.4}$$

As $\lambda_k(n)$, $1 \leq n \leq N$, are eigenvalues of $\mathbf{S}_k$, $1 \leq k \leq K$, we call $\mathbf{\Lambda}$ as the *joint spectrum* of commutative graph shifts $\mathbf{S}_1, \ldots, \mathbf{S}_K$ [20].

In this paper, we make the following assumption on the graph shifts $\mathbf{S}_1, \ldots, \mathbf{S}_K$ and their joint spectrum $\mathbf{\Lambda}$.

Assumption 2.1 Graph shifts $\mathbf{S}_1, \ldots, \mathbf{S}_K$ are real-valued, symmetric and commutative, and $\boldsymbol{\lambda}(n) \in \mathbb{R}^K$, $1 \leq n \leq N$, in the joint spectrum $\mathbf{\Lambda}$ in (2.4) are distinct.

2.2 Graph Fourier Transform and Graph Convolution

In this paper, we define the *graph Fourier transform* $\mathcal{F}\mathbf{x}$ of a graph signal $\mathbf{x} \in \mathbb{R}^V$ and the *inverse graph Fourier transform* $\mathcal{F}^{-1}\boldsymbol{\omega}$ of a vector $\boldsymbol{\omega} = [\omega(1), \ldots, \omega(N)]^T \in \mathbb{R}^N$

by

$$\mathcal{F}\mathbf{x} = \mathbf{U}^T\mathbf{x} = [\mathbf{u}_1^T\mathbf{x}, \ \dots \ , \mathbf{u}_N^T\mathbf{x}]^T \ \text{ and } \ \mathcal{F}^{-1}\boldsymbol{\omega} = \mathbf{U}\boldsymbol{\omega} = \sum_{n=1}^{N} \omega(n)\mathbf{u}_n, \tag{2.5}$$

where $\mathbf{U} = [\mathbf{u}_1, \ \dots \ , \mathbf{u}_N]$ is the orthogonal matrix in (2.3) to diagonalize the graph shifts $\mathbf{S}_1, \ \dots \ , \mathbf{S}_K$ simultaneously. The conventional definition of the graph Fourier transform on (un)directed graphs is based on one graph shift and a common selection of the graph shift is either the Laplacian matrix $\mathbf{L}$ or the symmetric normalized Laplacian matrix $\mathbf{L}^{\text{sym}}$ on the graph [9, 14, 20, 46, 64]. By (2.3), the Parseval identity holds for the graph Fourier transform $\mathcal{F}$ in (2.5),

$$\|\mathcal{F}\mathbf{x}\|_2 = \|\mathbf{x}\|_2 \ \text{ for all } \mathbf{x} \in \mathbb{R}^V. \tag{2.6}$$

We say that $\mathbf{H}$ is a *polynomial filter* of $\mathbf{S}_1, \dots, \mathbf{S}_K$ if

$$\mathbf{H} = h(\mathbf{S}_1, \ \dots \ , \mathbf{S}_K) = \sum h_{l_1,\dots,l_K} \mathbf{S}_1^{l_1} \cdots \mathbf{S}_K^{l_K} \tag{2.7}$$

for some multivariate polynomial $h(t_1, \ \dots \ , t_K) = \sum h_{l_1,\dots,l_K} \prod_{k=1}^{K} t_k^{l_k}$, where the sum is taken on a finite subset of $\mathbb{Z}_+^K$ [20, 28, 29, 46, 47, 53, 54, 60, 64]. We observe that the operation of a polynomial filter $\mathbf{H}$ of graph shifts $\mathbf{S}_k$, $1 \le k \le K$, becomes a multiplier $m(\mathbf{H})$ in the Fourier domain,

$$\mathcal{F}\mathbf{H}\mathbf{x} = m(\mathbf{H}) \odot (\mathcal{F}\mathbf{x}), \ \mathbf{x} \in \mathbb{R}^V, \tag{2.8}$$

where $\mathbf{a} \odot \mathbf{b}$ is the Hadamard product of two vectors $\mathbf{a}$ and $\mathbf{b} \in \mathbb{R}^N$. In particular, the multipliers associated with the graph shifts $\mathbf{S}_k$ are the diagonal vectors of the diagonal matrix $\mathbf{\Lambda}_k$, $1 \le k \le K$.

Given two graph signals $\mathbf{b}$ and $\mathbf{x}$, define their *convolution* $\mathbf{b} * \mathbf{x}$ by

$$\mathbf{b} * \mathbf{x} := \mathcal{F}^{-1}(\mathcal{F}\mathbf{b} \odot \mathcal{F}\mathbf{x}) = \sum_{n=1}^{N} (\mathbf{u}_n^T\mathbf{b})\mathbf{u}_n\mathbf{u}_n^T\mathbf{x}. \tag{2.9}$$

By (2.8) and (2.9), we see that the convolution associated with a graph signal $\mathbf{b}$ commutes with graph shifts $\mathbf{S}_k$, $1 \le k \le K$, i.e.,

$$\mathbf{S}_k(\mathbf{b} * \mathbf{x}) = (\mathbf{S}_k\mathbf{b}) * \mathbf{x}, \ \mathbf{x} \in \mathbb{R}^V.$$

With Assumption 2.1, it is shown in [20] that $\mathbf{H}$ is a polynomial filter if and only if it commutes with $\mathbf{S}_1, \dots, \mathbf{S}_K$, i.e., $\mathbf{H}\mathbf{S}_k = \mathbf{S}_k\mathbf{H}$, $1 \le k \le K$. Therefore the convolution operation associated with a graph signal $\mathbf{b}$ could be written as a polynomial filtering procedure,

$$\mathbf{b} * \mathbf{x} = h(\mathbf{S}_1, \ \dots \ , \mathbf{S}_K)\mathbf{x}, \ \mathbf{x} \in \mathbb{R}^V, \tag{2.10}$$

where h is a multivariate polynomial. In particular, we can show that (2.10) holds if and only if the polynomial h satisfies the following interpolation property

$$h(\boldsymbol{\lambda}(n)) = \mathbf{u}_n^T \mathbf{b}, \ 1 \le n \le N, \tag{2.11}$$

or equivalently

$$\mathbf{b} = h(\mathbf{S}_1, \ \dots, \mathbf{S}_K) \sum_{n=1}^{N} \mathbf{u}_n = h(\mathbf{S}_1, \ \dots, \mathbf{S}_K)\mathbf{U}\mathbf{1}, \tag{2.12}$$

where $\mathbf{1}$ is the column vector with all components taking value one.

Let $0 \le L \le N-1$. Denote the space of all multivariate polynomials of degree at most L by Π_L, and set

$$\mathcal{W}_L = \{\mathbf{b} = h(\mathbf{S}_1, \ \dots, \mathbf{S}_K)\mathbf{U}\mathbf{1} : \ h \in \Pi_L\}. \tag{2.13}$$

The spatial representation (2.10) of the convolution operation provides another approach to implement the convolution between graph signals $\mathbf{b}$ and $\mathbf{x}$ in the spatial domain. In particular, given a graph signal $\mathbf{b} \in \mathcal{W}_L$, we first evaluate the Fourier coefficients $\mathbf{u}_n^T\mathbf{b}$, $1 \le n \le N$, then find the multivariate polynomial h that takes values $\mathbf{u}_n^T\mathbf{b}$ at the spectrum $\boldsymbol{\lambda}(n)$, $1 \le n \le N$; and finally used the distributed algorithm to implement the polynomial filtering procedure in (2.10), see [20]. The total computational complexity to implement the distributed algorithm is about $O(L^K \times N)$.

2.3 Graph Convolutional Neural Networks

Let $\Omega \subset \mathbb{R}^V$ be a compact set of graph signal inputs of GCNNs. Due to the compactness of the set Ω, there exists a positive constant D_0 such that

$$\Omega \subset \{\mathbf{x} \in \mathbb{R}^V : \ \|\mathbf{x}\|_\infty \le D_0\}. \tag{2.14}$$

An illustrative example of the above compact domain is the set of all s-sparse graph signals whose norms are bounded by one,

$$\Omega_{\rm sp} = \{\mathbf{x} \in \mathbb{R}^V : \ \|\mathbf{x}\|_\infty \le 1, \ \|\mathbf{x}\|_0 \le s\}, \tag{2.15}$$

where we denote the standard ℓ^p-norm on the linear space of p-summable graph signals by $\|\cdot\|_p$, $1 \le p \le \infty$, and the number of nonzero entries of the vector $\mathbf{x}$ by $\|\mathbf{x}\|_0$. Taking $s = N$ in (2.15) leads to the unit ball with one as its radius and the origin as its center,

$$\Omega_{\rm ba} = \{\mathbf{x} \in \mathbb{R}^V : \ \|\mathbf{x}\|_\infty \le 1\} = [-1, 1]^V. \tag{2.16}$$

The uniform approximation property via shallow GCNNs on the above two illustrative domains is discussed in Remarks 4.4 and 4.5. In the classical neural network setting, a popular selection of the domain Ω is the unit cube $[0, 1]^N$ [3, 8, 17].

Let $\|\cdot\|$ be a norm on $\mathbb{R}^V$ **normalized** so that the ReLU activation function σ in GCNNs satisfy

$$\|\sigma(\mathbf{x})\| \le \|\mathbf{x}\| \quad \text{for all } \mathbf{x} \in \mathbb{R}^V. \tag{2.17}$$

Denote its dual norm by $\|\cdot\|_*$. Due to the norm equivalence on a finite-dimensional linear space, one may verify that the ReLU activation function σ has Lipschitz property on Ω with Lipschitz constant denoted by $\|\sigma\|_{\mathrm{Lip}}$,

$$\|\sigma(\mathbf{x}) - \sigma(\mathbf{x}')\| \le \|\sigma\|_{\mathrm{Lip}}\|\mathbf{x} - \mathbf{x}'\| \quad \text{for all } \mathbf{x}, \mathbf{x}' \in \mathbb{R}^N. \tag{2.18}$$

Our illustrative examples of the norm and its dual norm are the p-norm $\|\cdot\|_p$ and its dual q-norm $\|\cdot\|_q$, where $1/p + 1/q = 1$. In this case, the corresponding Lipschitz constant for the ReLU function σ satisfies $\|\sigma\|_{\mathrm{Lip}} = 1$.

Let $\mathcal{W}$ be a linear space of graph signals used for the convolution in GCNNs. For efficient learning of shallow GCNNs, our favorable assumption is that graph convolutions belong to the space $\mathcal{W}_L$ in (2.13) for some small $L \le N-1$. Consequently, graph convolution associated with a graph signal in $\mathcal{W}_L$ can be implemented by the polynomial filtering procedure in a distributed manner (and hence the shallow GCNNs could be learnt at the vertex level with limited coordination). In the standard CNN setting, a popular selection of graph convolutions is the family of $N \times N$ symmetric Toeplitz matrices with bandwidth L, where the shifting structure of Toeplitz matrices can be described by the circular graph.

For a graph signal $\mathbf{b} \in \mathcal{W}$, define a *convolution norm* $\|\mathbf{b}\|_{\mathrm{co}}$ such that

$$\|\mathbf{b} * \mathbf{x}\| \le \|\mathbf{b}\|_{\mathrm{co}} \quad \text{for all } \mathbf{x} \in \Omega. \tag{2.19}$$

To consider the Lipschitz property of functions in the graph Barron space in Corollary 3.4 and uniform approximation property in Theorem 4.6, we also require that the convolution norm satisfies the Lipschitz property with Lipschitz bounded by a multiple of the convolution norm,

$$\|\mathbf{b} * (\mathbf{x} - \mathbf{x}')\| \le D_1\|\mathbf{b}\|_{\mathrm{co}}\|\mathbf{x} - \mathbf{x}'\| \quad \text{for all } \mathbf{x}, \mathbf{x}' \in \Omega, \tag{2.20}$$

where D_1 is a positive constant. To consider the Rademacher complexity in Theorem 5.1 and learnability of functions in the graph Barron space in Theorem 5.4, we need the following Lipschitz property (2.20) for the convolution:

$$\Big\|\mathbf{b} * \Big(\sum_{i=1}^{S} \epsilon_i \mathbf{x}_i\Big)\Big\| \le D_2\|\mathbf{b}\|_{\mathrm{co}}\Big\|\sum_{i=1}^{S} \epsilon_i \mathbf{x}_i\Big\| \tag{2.21}$$

hold for all $\epsilon_i \in \{-1, 1\}$ and $\mathbf{x}_i \in \Omega$, $1 \le i \le S$ and all $S \ge 1$, where D_2 is a positive constant. We remark that the constants D_1 and D_2 in (2.20) and (2.21) always exist, since all norms on a finite-dimensional vector space are equivalent. However, the optimal values of these constants depend on the choice of the convolution norm $\|\cdot\|_{\mathrm{co}}$

in (2.19) and on the geometric structure of the domain Ω. Our illustrative example of the convolution norm $\|\mathbf{b}\|_{\mathrm{co}}$ of a graph signal $\mathbf{b} \in \mathcal{W}$ is

$$\|\mathbf{b}\|_{\mathrm{co}} = D_3 \|\mathbf{b}\|_{\mathrm{coop}},$$

where the constant D_3 is so chosen that

$$\|\mathbf{x}\| \le D_3, \quad \mathbf{x} \in \Omega, \tag{2.22}$$

and

$$\|\mathbf{b}\|_{\mathrm{coop}} = \sup_{\mathbf{0} \neq \mathbf{x} \in \mathbb{R}^V} \|\mathbf{b} * \mathbf{x}\| / \|\mathbf{x}\| \tag{2.23}$$

is the operator norm of the convolution $\mathbf{b}*$. For the above setting, the constants D_1 in (2.20) and D_2 in (2.21) are given by $D_1 = D_2 = D_3$.

If the convolution associated with the graph signal $\mathbf{b} \in \mathcal{W}_L$ can be represented by a polynomial filter $h(\mathbf{S}_1, \ldots, \mathbf{S}_K)$ in (2.12), a widely-used norm for the convolution is defined by an appropriate scaling of

$$\|\mathbf{b}\|_{\mathrm{cofi}} = \sum |h_{l_1,\ldots,l_K}| \prod_{k=1}^{K} \|\mathbf{S}_k\|^{l_k},$$

where $h(t_1, \ldots, t_K) = \sum h_{l_1,\ldots,l_K} \prod_{k=1}^{K} t_k^{l_k}$ and $\|\mathbf{S}_k\|$ is the operator norm of graph shifts $\mathbf{S}_k$, $1 \le k \le K$ [8, 15, 33, 70].

Barron space of functions on the unit cube $[0, 1]^N$ was introduced in [3], where it is shown that functions in Barron space are well approximated by the classical shallow neural networks. In this paper, we introduce a Barron space $\mathcal{B}$ of functions f of graph signals $\mathbf{x} \in \Omega$, and discuss its approximation property by some shallow GCNNs with M neurons at every vertex of the underlying graph $\mathcal{G}$, i.e.,

$$f(\mathbf{x}) \approx f_M(\mathbf{x}, \boldsymbol{\Theta}) := \frac{1}{M} \sum_{m=1}^{M} \mathbf{a}_m^T \sigma(\mathbf{b}_m * \mathbf{x} + \mathbf{c}_m), \quad \mathbf{x} \in \Omega$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$ and $\boldsymbol{\theta}_m = (\mathbf{a}_m, \mathbf{b}_m, \mathbf{c}_m) \in \mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V$, $1 \le m \le M$; see Sections 3, 4 and 5 for theoretical results and Section 6 for numerical demonstrations.

3 Barron Space on Graphs

Let $\mathcal{G} = (V, E)$ be a undirected graph of order $N \ge 2$, graph shifts $\mathbf{S}_1, \ldots, \mathbf{S}_K$ on the graph $\mathcal{G}$ satisfy Assumption 2.1, Ω be the domain for graph signal inputs of GCNNs, $\mathcal{W}$ be the linear space of graph signals used for the convolution in GCNNs, σ be the ReLU activation function in GCNNs, and the norm $\|\cdot\|$, its dual norm $\|\cdot\|_*$ and the convolution norm $\|\cdot\|_{\mathrm{co}}$ be as in Section 2.3. Our illustrative choices for the domain

Ω, the convolution norm $\|\cdot\|_{\rm co}$ and the vector norms $\|\cdot\|$ and $\|\cdot\|_*$, are

$$\Omega = \{\mathbf{x} \in \mathbb{R}^V : \|\mathbf{x}\|_\infty \le 1\},\ \|\cdot\| = \|\cdot\|_\infty,\ \|\cdot\|_* = \|\cdot\|_1 \text{ and } \|\mathbf{b}\|_{\rm co} = \|\mathbf{b}\|_{\rm coop}, \quad (3.1)$$

for which the constants D_0, D_1, D_2, D_3 in (2.14), (2.20) (2.21), (2.22) and the Lipschitz constant for the ReLU function σ are given by

$$D_0 = D_1 = D_2 = D_3 = 1 \ \text{ and } \ \|\sigma\|_{\rm Lip} = 1, \quad (3.2)$$

respectively. In this case, we also have

$$\|\mathbf{b}\|_{\rm co} \le \|\mathbf{b}\|_2 \|\mathbf{U}\|_{\mathcal{S}}^2 \le N\|\mathbf{b}\|_2, \quad (3.3)$$

where the Schur norm $\|\mathbf{U}\|_{\mathcal{S}}$ of the orthogonal matrix $\mathbf{U} = [u_n(v)]_{1\le n\le N, v\in V}$ in the definition of the GFT and convolution is defined by

$$\|\mathbf{U}\|_{\mathcal{S}} = \max\Big(\sup_{v\in V}\sum_{n=1}^N |u_n(v)|,\ \sup_{1\le n\le N}\sum_{v\in V}|u_n(v)|\Big).$$

Barron space of functions on the unit cube $[0,1]^N$ was introduced in [3] with the help of Fourier transform. In [2], Bach considered the space $\mathcal{F}_1$ of functions f with the following spatial representation

$$f(\mathbf{x}) = \int_{\mathcal{V}} \phi_z(\mathbf{x})\rho(dz),\ \mathbf{x} \in \Omega, \quad (3.4)$$

where $\phi_z, z \in \mathcal{V}$, is a family of basis functions (a.k.a neurons) and ρ is a signed Radon measure on $\mathcal{V}$ with finite total variation $|\rho|(\mathcal{V})$. In [17, 19], E, Ma and Wu introduced a Barron space of functions in (3.4) with ρ being a probability measure and $\phi_z(\mathbf{x}) = u\sigma(\mathbf{v}^T\mathbf{x} + w)$ being neurons, where $z = (u, \mathbf{v}, w) \in \mathbb{R}\times\mathbb{R}^N\times\mathbb{R}$. In this paper, we consider functions $f : \Omega \to \mathbb{R}$ on the domain Ω of graph signals that can be written as

$$f(\mathbf{x}) = \int_{\mathbb{R}^N\times\mathcal{W}\times\mathbb{R}^N} \mathbf{a}^T\sigma\left(\mathbf{b}*\mathbf{x}+\mathbf{c}\right)\rho\left(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}\right),\ \mathbf{x}\in\Omega, \quad (3.5)$$

where ρ is a probability measure on $\mathbb{R}^V\times\mathcal{W}\times\mathbb{R}^V$. We remark that functions in (3.5) have the spatial representation of the form (3.4) with neurons $\phi_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{a}^T\sigma(\mathbf{b}*\mathbf{x}+\mathbf{c})$ of GCNNs, where $\boldsymbol{\theta} = (\mathbf{a}, \mathbf{b}, \mathbf{c}) \in \mathbb{R}^V\times\mathcal{W}\times\mathbb{R}^V$.

In this section, we introduce graph Barron spaces $\mathcal{B}_r$, $1 \le r \le \infty$, of functions on the domain Ω with the spatial representation (3.5). Here for $1 \le r \le \infty$, let $\mathcal{B}_r := \mathcal{B}_r(\Omega, \mathcal{W})$ contain all functions f on the domain Ω with the spatial representation (3.5) such that $\mathbb{E}_\rho\big[\|\mathbf{a}\|_*^r(\|\mathbf{b}\|_{\rm co} + \|\mathbf{c}\|)^r\big] < \infty$ if $1 \le r < \infty$, and the support of the

probability measure ρ being bounded if $r = \infty$. Define the norm $\|f\|_{\mathcal{B}_r}$ of a function $f \in \mathcal{B}_r$ by

$$\|f\|_{\mathcal{B}_r} = \begin{cases} \inf\limits_{\rho\in\mathcal{P}_f} \left[\mathbb{E}_\rho\big(\|\mathbf{a}\|_*^r (\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|)^r\big)\right]^{1/r} & \text{if } 1 \le r < \infty \\ \inf\limits_{\rho\in\mathcal{P}_f} \max\limits_{(\mathbf{a},\mathbf{b},\mathbf{c})\in\mathrm{supp}(\rho)} \|\mathbf{a}\|_* (\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|) & \text{if } r = \infty, \end{cases} \tag{3.6}$$

where $\mathcal{P}_f$ is the collection of all possible probability measures ρ in the representation (3.5).

Reproducing kernel Banach spaces (RKBSs) accommodate a wider variety of norms and geometries, making them especially useful for modeling data with sparsity representations. This flexibility has led to their widespread use in neural networks, machine learning, sampling theory, sparse approximation, and functional analysis [5, 22, 38, 44, 62, 68, 72, 79, 80]. In Section 3.1, we show that the spaces $\mathcal{B}_r$ are RKBSs with norms independent of $1 \le r \le \infty$, and that functions in those spaces are Lipschitz continuous; see Theorem 3.1 and Corollary 3.4. Due to the norm independence, we denote the RKBSs $\mathcal{B}_r$, $1 \le r \le \infty$, by $\mathcal{B} := \mathcal{B}(\Omega, \mathcal{W})$ and define its norm by $\|\cdot\|_{\mathcal{B}}$, i.e.,

$$\|\cdot\|_{\mathcal{B}} = \|\cdot\|_{\mathcal{B}_r},\ 1 \le r \le \infty, \text{ and } \mathcal{B} = \{f :\ \|f\|_{\mathcal{B}} < \infty\}. \tag{3.7}$$

Following the terminology in [17], we refer to the RKBS $\mathcal{B}$ as the graph Barron space. In Section 3.1, we also include several remarks on the Barron space in the classical neural network setting, its connection to the graph Barron space, as well as variations of the graph Barron space arising from different choices of norms, activation functions, and neurons; see Remarks 3.5, 3.6, 3.7, 3.8 and 3.9.

Let

$$\mathbb{S} = \{\mathbf{a} \in \mathbb{R}^V :\ \|\mathbf{a}\|_* = 1\} \text{ and } \mathbb{T} = \{(\mathbf{b}, \mathbf{c}) \in \mathcal{W} \times \mathbb{R}^V :\ \|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\| = 1\} \tag{3.8}$$

be the unit spheres in $\mathbb{R}^V$ and $\mathcal{W} \times \mathbb{R}^V$ respectively, and $\widehat{P}$ be the set of all probability measures on $\mathbb{S} \times \mathbb{T}$. A crucial step in the proof of Theorem 3.1 is the following spatial representation of functions f in the graph Barron space $\mathcal{B}$ for some probability measure $\hat{\rho}$ on $\mathbb{S} \times \mathbb{T}$,

$$f(\mathbf{x}) = \|f\|_{\mathcal{B}} \int_{\mathbb{S}\times\mathbb{T}} \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}) \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}), \tag{3.9}$$

see Lemma 3.3.

Reproducing kernel Hilbert spaces (RKHSs) have been widely adopted in kernel-based learning for function estimation, offering dimension-independent error rates; see [24, 51, 55, 65] and references therein. Their kernels are typically chosen to encode specific notions of similarity between input data, which can lead to significant computational savings. Moreover, in contrast to RKBSs, the orthogonal projections and spectral properties inherent to RKHSs confer additional advantages for analysis and computation. The graph Barron space $\mathcal{B}$ is a reproducing kernel Banach space for which an explicit expression for the reproducing kernel is not available. This raises the question of whether the graph Barron space can be expressed as a union of RKHSs

with explicit reproducing kernels and an associated norm equivalence. We provide an affirmative answer to this question in Sections 3.2 and 3.3.

In Section 3.2, we introduce RKHSs $\mathcal{H}_{\hat{\rho}}$, $\hat{\rho} \in \widehat{P}$, whose kernel functions $K_{\hat{\rho}}$ are defined by

$$K_{\hat{\rho}}(\mathbf{x}, \mathbf{x}') = \int_{\mathbb{S}\times\mathbb{T}} \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}) \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x}' + \mathbf{c}) \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}), \ \mathbf{x}, \mathbf{x}' \in \Omega. \quad (3.10)$$

In Theorem 3.10, we show that any function g in the RKHS $\mathcal{H}_{\hat{\rho}}$ admits the following integral representation, together with a corresponding norm estimate:

$$g(\mathbf{x}) = \int_{\mathbb{S}\times\mathbb{T}} \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}) \eta(\mathbf{a}, \mathbf{b}, \mathbf{c}) \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \quad (3.11)$$

for some function $\eta \in L^2_{\hat{\rho}} := L^2_{\hat{\rho}}(\mathbb{S} \times \mathbb{T})$, the space of square-integrable functions on $\mathbb{S} \times \mathbb{T}$ with respect to the probability measure $\hat{\rho}$. More importantly, in Section 3.3, we show that the graph Barron space $\mathcal{B}$ is the union of the RKHSs $\mathcal{H}_{\hat{\rho}}$, $\hat{\rho} \in \widehat{P}$, with an associated norm equivalence; see Theorem 3.12. We remark that in the classical neural network setting, spaces $\mathcal{F}_1$ and $\mathcal{F}_2$, analogous to the RKHSs $\mathcal{H}_{\hat{\rho}}$, $\hat{\rho} \in P$ in our GCNNs, are introduced in [2, 17]. In that setting, functions in $\mathcal{F}_1$ and $\mathcal{F}_2$ admit representations similar to (3.46), with η required to be integrable or square-integrable on some compact set, respectively. We also remark that, in the classical neural network setting, the Barron space can be expressed as a union of a family of RKHSs; see [17, Proposition 3]. In contrast to our GCNN setting in Theorem 3.12, the associated norm equivalence is not addressed in [17].

3.1 Barron Spaces and Reproducing Kernel Banach Spaces

Let $C(\Omega)$ be the Banach space of continuous functions on the domain Ω with the uniform norm defined by

$$\|f\|_\infty = \sup_{\mathbf{x}\in\Omega} |f(\mathbf{x})|, \quad f \in C(\Omega).$$

In the following theorem, we show that the normed vector space $\mathcal{B}_r := \mathcal{B}_r(\Omega, \mathcal{W})$ in (3.6) is a reproducing kernel Banach subspace of $C(\Omega)$, with norms independent of $1 \le r \le \infty$.

Theorem 3.1 *Let graph shifts $\mathbf{S}_1, \dots, \mathbf{S}_K$ on the graph $\mathcal{G}$ satisfy Assumption 2.1, Ω be the domain for graph signal inputs of GCNNs, $\mathcal{W}$ be the linear space of graph signals used for the convolution in GCNNs, σ be the ReLU activation function in GCNNs, and the norm $\|\cdot\|$, its dual norm $\|\cdot\|_*$ and the convolution norm $\|\cdot\|_{\mathrm{co}}$ be as in Section 2.3. Then $\mathcal{B}_r(\Omega, \mathcal{W})$, $1 \le r \le \infty$, in (3.6) are the same RKBS. Moreover,*

$$\|f\|_\infty \le \|f\|_{\mathcal{B}_r}, \quad (3.12)$$

and

$$\|f\|_{\mathcal{B}_\infty} = \|f\|_{\mathcal{B}_1} = \|f\|_{\mathcal{B}_r} \tag{3.13}$$

hold for all $f \in \mathcal{B}_r(\Omega, \mathcal{W})$, $1 \le r \le \infty$.

To prove Theorem 3.1, we first show that $\|\cdot\|_{\mathcal{B}_1}$ in (3.6) defines a norm on the Barron space $\mathcal{B}_1$.

Lemma 3.2 *Let* $\|\cdot\|_{\mathcal{B}_1}$ *be as in (3.6). Then*

(i) $f = 0$ *if and only if* $\|f\|_{\mathcal{B}_1} = 0$.
(ii) $\|\alpha f\|_{\mathcal{B}_1} = |\alpha|\|f\|_{\mathcal{B}_1}$ *for all* $f \in \mathcal{B}_1$ *and* $\alpha \in \mathbb{R}$.
(iii) $\|f + g\|_{\mathcal{B}_1} \le \|f\|_{\mathcal{B}_1} + \|g\|_{\mathcal{B}_1}$ *for all* $f, g \in \mathcal{B}_1$.

Proof (i) Taking the Dirac measure δ_0 at the origin as the probability measure in (3.5) gives a representation for the zero function. This shows that $\|f\|_{\mathcal{B}_1} = 0$ for the zero function $f = 0$. Conversely, given $f \in \mathcal{B}_1$ with $\|f\|_{\mathcal{B}_1} = 0$, there exists a probability measure ρ for any $\epsilon > 0$ such that (3.5) holds and $\mathbb{E}_\rho(\|\mathbf{a}\|_*(\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|)) \le \epsilon$. Therefore for any $\mathbf{x} \in \Omega$, we have

$$\begin{aligned} |f(\mathbf{x})| &\le \int_{\mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V} \|\mathbf{a}\|_*(\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|)\rho(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \\ &= \mathbb{E}_\rho(\|\mathbf{a}\|_*(\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|)) \le \epsilon. \end{aligned}$$

As $\epsilon > 0$ is arbitrarily chosen, we conclude that f must be the zero function on the domain Ω. This proves the conclusion (i).

(ii) Clearly it suffices to show that

$$\|\alpha f\|_{\mathcal{B}_1} \le |\alpha|\|f\|_{\mathcal{B}_1} \tag{3.14}$$

for all $0 \ne \alpha \in \mathbb{R}$ and $f \in \mathcal{B}_1$. Take arbitrary $\epsilon > 0$ and let ρ be a probability measure in $\mathcal{P}_f$ such that

$$\mathbb{E}_\rho(\|\mathbf{a}\|_*(\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|)) \le \|f\|_{\mathcal{B}_1} + \epsilon. \tag{3.15}$$

Define a new probability measure $\tilde{\rho}(A) = \rho(\tilde{A})$ for any Borel set A, where $\tilde{A} = \{(\mathbf{a}, \mathbf{b}, \mathbf{c}) | (\alpha\mathbf{a}, \mathbf{b}, \mathbf{c}) \in A\}$. Then one may verify that

$$\begin{aligned} \alpha f(\mathbf{x}) &= \int_{(\mathbf{a},\mathbf{b},\mathbf{c}) \in \mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V} (\alpha\mathbf{a})^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c})\rho(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \\ &= \int_{(\tilde{\mathbf{a}},\mathbf{b},\mathbf{c}) \in \mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V} \tilde{\mathbf{a}}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c})\tilde{\rho}(d\tilde{\mathbf{a}}, d\mathbf{b}, d\mathbf{c}) \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}_{\tilde{\rho}}\big(\|\mathbf{a}\|_*(\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|)\big) &= \int_{(\mathbf{a},\mathbf{b},\mathbf{c}) \in \mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V} \|\mathbf{a}\|_*(\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|)\tilde{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \\ &= \int \|\alpha\tilde{\mathbf{a}}\|_*(\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|)\rho(d\tilde{\mathbf{a}}, d\mathbf{b}, d\mathbf{c}) \\ &= |\alpha|\mathbb{E}_\rho(\|\tilde{\mathbf{a}}\|_*(\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|)) \le |\alpha|\|f\|_{\mathcal{B}_1} + |\alpha|\epsilon. \end{aligned}$$

Then the desired estimate (3.14) follows from the above estimate and the arbitrary selection of $\epsilon > 0$.

(iii) By the second conclusion, it suffices to prove that

$$\|\alpha f_1 + (1-\alpha) f_2\|_{\mathcal{B}_1} \le \alpha \|f_1\|_{\mathcal{B}_1} + (1-\alpha)\|f_2\|_{\mathcal{B}_1}, \ f_1, f_2 \in \mathcal{B}_1 \tag{3.16}$$

where $0 \le \alpha \le 1$. Take arbitrary $\epsilon > 0$ and let $\rho_1 \in \mathcal{P}_{f_1}$ and $\rho_2 \in \mathcal{P}_{f_2}$ be two probability measures so that

$$\mathbb{E}_{\rho_l}(\|\mathbf{a}\|_*(\|\mathbf{b}\|_{\rm co} + \|\mathbf{c}\|)) \le \|f_l\|_{\mathcal{B}_1} + \epsilon, \ l = 1, 2. \tag{3.17}$$

Define $\rho = \alpha\rho_1 + (1-\alpha)\rho_2$ and set $f = \alpha f_1 + (1-\alpha) f_2$. Then one may verify that ρ is a probability measure in $\mathcal{P}_f$ and

$$\mathbb{E}_{\rho}(\|\mathbf{a}\|_*(\|\mathbf{b}\|_{\rm co} + \|\mathbf{c}\|)) \le \alpha\|f_1\|_{\mathcal{B}_1} + (1-\alpha)\|f_2\|_{\mathcal{B}_1} + \epsilon.$$

This together with the arbitrary selection of $\epsilon > 0$ proves (3.16). □

To prove Theorem 3.1, we next show that the probability measure ρ in the representation (3.5) of any function f in $\mathcal{B}_1$ could be selected to be supported on the dilated unit sphere.

Lemma 3.3 *Let $\mathbb{S}$ and $\mathbb{T}$ be as in (3.8). Then for any $f \in \mathcal{B}_1$, there exists a probability measure $\hat{\rho}$ supported on $\mathbb{S} \times \mathbb{T}$ such that*

$$f(\mathbf{x}) = \|f\|_{\mathcal{B}_1} \int_{\mathbb{S}\times\mathbb{T}} \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}) \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}), \ \mathbf{x} \in \Omega. \tag{3.18}$$

Proof The conclusion is obvious for the zero function. Now we assume that $f \neq 0$. By (3.6), there exist probability measures $\rho_n \in \mathcal{P}_f, n \ge 1$, on $\mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V$ such that

$$0 < \|f\|_{\mathcal{B}_1} \le A_n := \mathbb{E}_{\rho_n}\big(\|\mathbf{a}\|_*(\|\mathbf{b}\|_{\rm co} + \|\mathbf{c}\|)\big) \le \|f\|_{\mathcal{B}_1} + 2^{-n}, \ n \ge 1. \tag{3.19}$$

Set

$$\mathbf{O} = \Big\{(\mathbf{a}, \mathbf{b}, \mathbf{c}) \in \mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V : \text{ either } \mathbf{a} = \mathbf{0} \text{ or } (\mathbf{b}, \mathbf{c}) = \mathbf{0}\Big\}.$$

Without loss of generality, we assume that

$$\rho_n(\mathbf{O}) = 0. \tag{3.20}$$

Otherwise, replacing ρ_n by another probability measure $\tilde{\rho}_n \in \mathbb{P}_f$ for which

$$\tilde{\rho}_n(\mathbf{O}) = 0, \tag{3.21}$$

and

$$\mathbb{E}_{\tilde{\rho}_n}\big(\|\mathbf{a}\|_*(\|\mathbf{b}\|_{\rm co} + \|\mathbf{c}\|)\big) = A_n, \tag{3.22}$$

where for a Borel set $E \subset \mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V$, we define

$$E' = \{(\mathbf{a}, \mathbf{b}, \mathbf{c}) \in \mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V : ((1 - \rho_n(\mathbf{O}))^{-1}\mathbf{a}, \mathbf{b}, \mathbf{c}) \in E\}$$

and

$$\tilde{\rho}_n(E) = (1 - \rho_n(\mathbf{O}))^{-1}\rho_n(E'\backslash\mathbf{O}).$$

The measure $\tilde{\rho}_n$ is well-defined as $0 < \rho_n(\mathbf{O}) < 1$ by (3.2) and the assumptions on f and ρ_n, provided that $\rho_n(\mathbf{O}) \neq 0$. Clearly, from the definition of the measure $\tilde{\rho}_n$, we see that $\tilde{\rho}_n$ is a probability measure satisfying (3.21). Moreover, $\tilde{\rho}_n$ belongs to $\mathcal{P}_f$ and satisfies (3.22), because

$$\begin{aligned} f(\mathbf{x}) &= \int_{(\mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V)\backslash\mathbf{O}} \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}) \rho_n(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \\ &= (1 - \rho_n(O)) \int_{(\mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V)\backslash\mathbf{O}} \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}) \tilde{\rho}_n(d((1 - \rho_n(O))\mathbf{a}), d\mathbf{b}, d\mathbf{c}) \\ &= \int_{\mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V} \tilde{\mathbf{a}}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}) \tilde{\rho}_n(d\tilde{\mathbf{a}}, d\mathbf{b}, d\mathbf{c}) \end{aligned}$$

by (3.21) and the construction of the measure $\tilde{\rho}_n$, and

$$\begin{aligned} &\mathbb{E}_{\tilde{\rho}_n}\big(\|\mathbf{a}\|_*(\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|)\big) \\ &= (1 - \rho_n(\mathbf{O})) \int_{(\mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V)\backslash\mathbf{O}} \|\mathbf{a}\|_*(\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|) \tilde{\rho}_n(d((1 - \rho_n(\mathbf{O}))\mathbf{a}), d\mathbf{b}, d\mathbf{c}) \\ &= \int_{(\mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V)\backslash\mathbf{O}} \|\mathbf{a}\|_*(\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|) \rho_n(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) = \mathbb{E}_{\rho_n}\big(\|\mathbf{a}\|_*(\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|)\big). \end{aligned}$$

For $n \geq 1$, define the measure $\hat{\rho}_n(E)$ of a Borel measurable subset $E \subset \mathbb{S} \times \mathbb{T}$ by

$$\hat{\rho}_n(E) = A_n^{-1}\mathbb{E}_{\rho_n}\big(\|\mathbf{a}\|_*(\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|)\chi_{\hat{E}}(\mathbf{a}, \mathbf{b}, \mathbf{c})\big) \tag{3.23}$$

where

$$\hat{E} = \left\{(\mathbf{a}, \mathbf{b}, \mathbf{c}) \in \mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V : \left(\frac{\mathbf{a}}{\|\mathbf{a}\|_*}, \frac{\mathbf{b}}{\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|}, \frac{\mathbf{c}}{\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|}\right) \in E\right\},$$

and $\chi_{\hat{E}}$ is the characteristic function on the set $\hat{E}$. By (3.20), (3.23) and the assumption $\rho_n \in \mathcal{P}_f$, we can verify that $\hat{\rho}_n$, $n \geq 1$, are probability measures on $\mathbb{S} \times \mathbb{T}$, and

$$\begin{aligned} f(\mathbf{x}) &= \int_{\mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V} \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}) \rho_n(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \\ &= \int_{\mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V} \|\mathbf{a}\|_* (\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|) \\ &\quad \times \Big(\frac{\mathbf{a}}{\|\mathbf{a}\|_*}\Big)^T \sigma\Big(\frac{\mathbf{b}}{\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|} * \mathbf{x} + \frac{\mathbf{c}}{\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|}\Big) \rho_n(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \\ &= A_n \int_{\mathbb{S} \times \mathbb{T}} \hat{\mathbf{a}}^T \sigma\big(\hat{\mathbf{b}} * \mathbf{x} + \hat{\mathbf{c}}\big) \hat{\rho}_n(d\hat{\mathbf{a}}, d\hat{\mathbf{b}}, d\hat{\mathbf{c}}). \end{aligned} \tag{3.24}$$

Recall that $\hat{\rho}_n$, $n \geq 1$, is a sequence of probability measures on the compact set $\mathbb{S} \times \mathbb{T}$. Then by Prokhorov theorem [50], without loss of generality, we assume that $\hat{\rho}_n$, $n \geq 1$, converges weakly to a probability measure $\hat{\rho}$ on $\mathbb{S} \times \mathbb{T}$,

$$\lim_{n \to \infty} \hat{\rho}_n = \hat{\rho} \ \text{ weakly}, \tag{3.25}$$

otherwise replacing the sequence by a weakly convergent subsequence. For any $\mathbf{x} \in \Omega$, the function $\mathbf{a}^T(\mathbf{b} * \mathbf{x} + \mathbf{c})$ is continuous with respect to $(\mathbf{a}, \mathbf{b}, \mathbf{c}) \in \mathbb{S} \times \mathbb{T}$ and is bounded by one. Therefore the desired conclusion (3.18) follows from (3.19), (3.24) and (3.25). □

Now we are ready to prove Theorem 3.1.

Proof of Theorem 3.1 First we prove that $\mathcal{B}_1$ is a Banach space. By Lemma 3.2, it suffices to prove every Cauchy sequence f_n, $n \geq 1$, in $\mathcal{B}_1$ converges to some function in $\mathcal{B}_1$. In particular, without loss of generality, we may assume that

$$\|f_{n+1} - f_n\|_{\mathcal{B}_1} \leq 2^{-n}, \ n \geq 1, \tag{3.26}$$

otherwise replacing it by one of its subsequences satisfying (3.26).

By Lemma 3.3, there exist probability measures $\hat{\rho}_n$, $n \geq 1$, on $\mathbb{S} \times \mathbb{T}$ such that

$$f_1(\mathbf{x}) = \|f_1\|_{\mathcal{B}_1} \int_{\mathbb{S} \times \mathbb{T}} \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}) \hat{\rho}_1(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \tag{3.27}$$

and

$$f_n(\mathbf{x}) - f_{n-1}(\mathbf{x}) = \|f_n - f_{n-1}\|_{\mathcal{B}_1} \int_{\mathbb{S} \times \mathbb{T}} \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}) \hat{\rho}_n(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}), \ \mathbf{x} \in \Omega \tag{3.28}$$

for all $n \geq 2$. Define

$$f(\mathbf{x}) = A \int_{\mathbb{S} \times \mathbb{T}} \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}) \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}), \ \mathbf{x} \in \Omega, \tag{3.29}$$

where $A = \|f_1\|_{\mathcal{B}_1} + \sum_{n=2}^{\infty} \|f_n - f_{n-1}\|_{\mathcal{B}_1} < \infty$ and the probability measure $\hat{\rho}$ on $\mathbb{S} \times \mathbb{T}$ is given by

$$\hat{\rho} = A^{-1}\Big(\|f_1\|_{\mathcal{B}_1}\hat{\rho}_1 + \sum_{n=2}^{\infty} \|f_n - f_{n-1}\|_{\mathcal{B}_1}\hat{\rho}_n\Big). \tag{3.30}$$

Dilate the measure $\hat{\rho}$ on $\mathbb{S} \times \mathbb{T}$ to a probability measure on $(A\mathbb{S}) \times \mathbb{T}$ and then extend to a probability measure on $\mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V$ with support on $(A\mathbb{S}) \times \mathbb{T}$. Denote the dilated extension measure by $\tilde{\rho}$. By (3.29) and (3.30), the dilated extension measure $\tilde{\rho}$ is a probability measure on $\mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V$ satisfying (3.5), i.e., $\tilde{\rho} \in \mathcal{P}_f$. Again from (3.29) and (3.30) we obtain

$$\|f\|_{\mathcal{B}_1} \le \mathbb{E}_{\tilde{\rho}}(\|\mathbf{a}\|_*(\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|)) = A,$$

which proves that $f \in \mathcal{B}_1$.

Extend the probability measure $\hat{\rho}_m, m \ge 1$ on $\mathbb{S} \times \mathbb{T}$ to probability measures $\tilde{\rho}_m$ on $\mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V$ with support on $\mathbb{S} \times \mathbb{T}$. We observe that

$$\begin{aligned}
\|f_n - f\|_{\mathcal{B}_1} &= \Big\| \sum_{m=n+1}^{\infty} \|f_m - f_{m-1}\|_{\mathcal{B}_1} \int_{\mathbb{S}\times\mathbb{T}} \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}) \hat{\rho}_m(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \Big\|_{\mathcal{B}_1} \\
&\le \sum_{m=n+1}^{\infty} \|f_m - f_{m-1}\|_{\mathcal{B}_1} \Big\| \int_{\mathbb{S}\times\mathbb{T}} \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}) \hat{\rho}_m(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \Big\|_{\mathcal{B}_1} \\
&\le \sum_{m=n+1}^{\infty} \|f_m - f_{m-1}\|_{\mathcal{B}_1} \mathbb{E}_{\tilde{\rho}_m}(\|\mathbf{a}\|_*(\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|)) \le 2^{-n+1},
\end{aligned}$$

where the equality holds by (3.27), (3.28) and (3.29), and the first, second and third inequality follows from Lemma 3.2, the definition of Barron norm and (3.26) respectively. Therefore $f_n, n \ge 1$, converges to $f \in \mathcal{B}_1$ and hence $\mathcal{B}_1$ is a Banach space.

By (2.17), (2.19) and (3.18), we have

$$\begin{aligned}
|f(\mathbf{x})| &\le \|f\|_{\mathcal{B}_1} \int_{\mathbb{S}\times\mathbb{T}} |\mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c})| \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \\
&\le \|f\|_{\mathcal{B}_1} \int_{\mathbb{S}\times\mathbb{T}} \|\mathbf{a}\|_* \|\mathbf{b} * \mathbf{x} + \mathbf{c}\| \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \\
&\le \|f\|_{\mathcal{B}_1} \int_{\mathbb{S}\times\mathbb{T}} \|\mathbf{a}\|_*(\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|) \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) = \|f\|_{\mathcal{B}_1}.
\end{aligned}$$

This proves the reproducing kernel property (3.12) for the Banach space $\mathcal{B}_1$.

Applying Hölder inequality, we have

$$\|f\|_{\mathcal{B}_1} \le \|f\|_{\mathcal{B}_r} \le \|f\|_{\mathcal{B}_\infty} \text{ for all } f \in \mathcal{B}_\infty \text{ and } 1 \le r \le \infty. \tag{3.31}$$

Therefore the proof of the norm equivalence in (3.13) reduces to establishing

$$\|f\|_{\mathcal{B}_\infty} \le \|f\|_{\mathcal{B}_1} \text{ for all } f \in \mathcal{B}_1. \tag{3.32}$$

By Lemma 3.3, there exists a probability measure $\hat{\rho}$ on $\mathbb{S} \times \mathbb{T}$ such that

$$f(\mathbf{x}) = \|f\|_{\mathcal{B}_1} \int_{\mathbb{S}\times\mathbb{T}} \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}) \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}). \tag{3.33}$$

Dilate the measure $\hat{\rho}$ on $\mathbb{S} \times \mathbb{T}$ to a probability measure on $(\|f\|_{\mathcal{B}_1}\mathbb{S}) \times \mathbb{T}$ and then extend to a probability measure $\tilde{\rho}$ on $\mathbb{R}^N \times \mathcal{W} \times \mathbb{R}^N$ with support on $(\|f\|_{\mathcal{B}_1}\mathbb{S}) \times \mathbb{T}$. Then one may verify that $\tilde{\rho} \in \mathcal{P}_f$ and

$$\|f\|_{\mathcal{B}_\infty} \le \sup_{(\mathbf{a},\mathbf{b},\mathbf{c})\in \text{supp}\tilde{\rho}} \|\mathbf{a}\|_*(\|\mathbf{b}\|_{\text{co}} + \|\mathbf{c}\|) = \|f\|_{\mathcal{B}_1}.$$

This proves (3.32). Hence the desired conclusion that $\mathcal{B}_r$, $1 \le r \le \infty$, are Banach spaces independent on $1 \le r \le \infty$. □

Under the additional assumption that the ReLU function σ and the convolution norm satisfy (2.18) and (2.20) respectively, for any $f \in \mathcal{B}$ and $\mathbf{x}, \mathbf{x}' \in \Omega$, we obtain from (3.9) that

$$\begin{aligned} |f(\mathbf{x}) - f(\mathbf{x}')| &\le \|f\|_{\mathcal{B}} \int_{\mathbb{S}\times\mathbb{T}} \|\mathbf{a}\|_* \|\sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}) - \sigma(\mathbf{b} * \mathbf{x}' + \mathbf{c})\| \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \\ &\le \|f\|_{\mathcal{B}} \int_{\mathbb{S}\times\mathbb{T}} \|\sigma\|_{\text{Lip}} \|\mathbf{b} * (\mathbf{x} - \mathbf{x}')\| \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \\ &\le D_1 \|\sigma\|_{\text{Lip}} \|f\|_{\mathcal{B}} \|\mathbf{x} - \mathbf{x}'\| \quad \text{for all } \mathbf{x}, \mathbf{x}' \in \Omega. \end{aligned} \tag{3.34}$$

Therefore functions in the graph Barron space $\mathcal{B}$ have Lipschitz property, cf. [3] and [19, Theorem 3.3].

Corollary 3.4 *Let $\mathcal{B}$ be the graph Barron space of functions on the domain Ω given in (3.7). If the ReLU activation function σ satisfies (2.17) and (2.18), and if the convolution norm $\|\cdot\|_{\text{co}}$ satisfies (2.19) and (2.20), then any function f in the graph Barron space $\mathcal{B}$ has the Lipschitz property with Lispchitz constant bounded by $D_1\|\sigma\|_{\text{Lip}}\|f\|_{\mathcal{B}}$, where $\|\sigma\|_{\text{Lip}}$ and D_1 are the constants in (2.18) and (2.20) respectively.*

We conclude this subsection with several remarks on the Barron space in the classical neural network setting, its connection to the graph Barron space, variations of the graph Barron space with different norms, activation functions and neuron selection.

Remark 3.5 In [17], E, Ma and Wu introduce Barron spaces $\mathcal{B}_{r,\text{EMW}}$, $1 \le r \le \infty$, of functions f on a compact domain $\Omega \subset \mathbb{R}^N$ that admit the following representation

$$f(\mathbf{x}) = \int_{\mathbb{R}\times\mathbb{R}^N\times\mathbb{R}} u\sigma(\mathbf{v}^T\mathbf{x} + w)\tilde{\rho}(du, d\mathbf{v}, dw), \ \mathbf{x} \in \Omega, \tag{3.35}$$

for some probability measure $\tilde{\rho}$ on a Borel σ-algebra of $\mathbb{R}^{N+2}$, and define

$$\|f\|_{\mathcal{B}_{r,\mathrm{EMW}}} = \inf\Big(\int_{\mathbb{R}\times\mathbb{R}^N\times\mathbb{R}} |u|^r(\|\mathbf{v}\|_1+|w|)^r\,\tilde{\rho}(du,d\mathbf{v},dw)\Big)^{1/r} \tag{3.36}$$

for $1 \le r < \infty$ with standard modification for $r = \infty$, where the infimum in (3.36) is taken in **all** probability measures $\tilde{\rho}$ such that f has the representation (3.35). It is proved in [17] that $\|f\|_{\mathcal{B}_{r,\mathrm{EMW}}} = \|f\|_{\mathcal{B}_{1,\mathrm{EMW}}} =: \|f\|_{\mathcal{B}_{\mathrm{EMW}}}$ for all $1 \le r \le \infty$. Following the argument used in the proof of Theorem 3.1, one may show that $\|\cdot\|_{\mathcal{B}_{\mathrm{EMW}}}$ defines a norm and the Barron space

$$\mathcal{B}_{\mathrm{EMW}} = \big\{f:\ \|f\|_{\mathcal{B}_{\mathrm{EMW}}} < \infty\big\} \tag{3.37}$$

of functions f on the domain $\Omega \subset \mathbb{R}^N$ is a Banach space, which are not mentioned and proved explicitly in [17].

Remark 3.6 Let $\mathbf{U} = [\mathbf{u}_1,\ldots,\mathbf{u}_N]$ be as in (2.3), and write $\mathbf{u}_n = [u_n(i)]_{i\in V}$, $1 \le n \le N$, and $\mathbf{a} = [a(i)]_{i\in V}$ and $\mathbf{c} = [c(i)]_{i\in V}$ for $\mathbf{a},\mathbf{c} \in \mathbb{R}^V$. For any function f represented by (3.5), we have

$$\begin{aligned} f(\mathbf{x}) &= \sum_{i\in V}\int_{\mathbb{R}^V\times\mathcal{W}\times\mathbb{R}^V} a(i)\sigma\Big(\sum_{n=1}^N(\mathbf{b}^T\mathbf{u}_n)\mathbf{u}_n(i)\mathbf{u}_n^T\mathbf{x}+c(i)\Big)\rho(d\mathbf{a},d\mathbf{b},d\mathbf{c}) \\ &= \sum_{i\in V}\int_{(u_i,\mathbf{v}_i,w_i)\in\mathbb{R}\times\mathbb{R}^V\times\mathbb{R}} u_i\sigma(\mathbf{v}_i^T\mathbf{x}+w_i)\rho_i(du_i,d\mathbf{v}_i,dw_i) \\ &= \int_{(u,\mathbf{v},w)\in\mathbb{R}\times\mathbb{R}^V\times\mathbb{R}} u\sigma(\mathbf{v}^T\mathbf{x}+w)\tilde{\rho}(du,d\mathbf{v},dw),\ \ \mathbf{x}\in\Omega, \end{aligned} \tag{3.38}$$

where for a Borel set $E \subset \mathbb{R}\times\mathbb{R}^V\times\mathbb{R}$, we define $\rho_i(E) = \rho(\widetilde{E}_m)$ with

$$\widetilde{E}_i = \Big\{(\mathbf{a},\mathbf{b},\mathbf{c}) \in \mathbb{R}^N\times\mathcal{W}\times\mathbb{R}^N:\ \Big(a(i),\sum_{n=1}^N(\mathbf{b}^T\mathbf{u}_n)\mathbf{u}_n(i)\mathbf{u}_n,c(i)\Big)\in E\Big\}$$

for $1 \le i \le N$, and set $\tilde{\rho}(E) = N^{-1}\sum_{i\in V}\rho_i(E')$ with

$$E' = \{(u,\mathbf{v},w) \in \mathbb{R}\times\mathbb{R}^N\times\mathbb{R}:\ (N^{-1}u,\mathbf{v},w)\in E\}.$$

Therefore a graph function f represented by (3.5), **without** the underlying graph structure into consideration, could be treated as a function on a domain of the Euclidean space $\mathbb{R}^N$ that admits the representation (3.35) with the probability measure $\tilde{\rho}$ obtained from some probability measure ρ on $\mathbb{R}^V\times\mathcal{W}\times\mathbb{R}^V$. Moreover, it follows (3.38) and equivalence of different norms on a finite-dimensional space, there exists a positive constant $C(N,\mathbf{U})$ depending on the order N of the graph $\mathcal{G}$ and the orthogonal matrix

U in (2.3) such that

$$\begin{aligned}\|f\|_{\mathcal{B}_{\mathrm{EMW}}} &\le \inf_{\tilde{\rho} \text{ in } (3.38)} \int_{\mathbb{R}\times\mathbb{R}^N\times\mathbb{R}} |u|(\|\mathbf{v}\|_1 + |w|)\tilde{\rho}(du, d\mathbf{v}, dw) \\ &\le C(N, \mathbf{U}) \inf_{\rho \text{ in } (3.5)} \int_{\mathbb{R}^V\times\mathcal{W}\times\mathbb{R}^V} \|\mathbf{a}\|_* (\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|)\rho(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \\ &= C(N, \mathbf{U})\|f\|_{\mathcal{B}}. \end{aligned} \tag{3.39}$$

Therefore functions in the graph Barron space $\mathcal{B}$ belongs to the Barron space $\mathcal{B}_{\mathrm{EMW}}$ proposed in [17]. However, we believe that not every function in the Barron space $\mathcal{B}_{\mathrm{EMW}}$ admits the representation (3.5) involving graph convolution and belongs to the proposed graph Barron space $\mathcal{B}$. Hence the graph Barron space $\mathcal{B}$ is a **true** set of the Barron space $\mathcal{B}_{\mathrm{EMW}}$ and some function in the Barron space $\mathcal{B}_{\mathrm{EMW}}$ may not be well approximated by shallow GCNNs.

Remark 3.7 Due to the equivalence of different norms in a finite-dimensional space (even if the equivalence constant could be very large), the graph Barron space $\mathcal{B}$ in (3.7) does not depend on the selection of norms on $\mathbf{a}, \mathbf{c} \in \mathbb{R}^V$ and $\mathbf{b} \in \mathcal{W}$; however, the Barron norm $\|\cdot\|_{\mathcal{B}}$ **does**. Selecting appropriate norms is crucial for obtaining neat estimations that are independent of the graph order N or other important parameters used in GCNNs, and determining the number M of neurons needed at each vertex for efficient learning of shallow GCNNs; see (3.12), (3.57), (4.10), (4.11), (4.15) and (4.16). In our definition, we select norms such that

$$|\mathbf{a}^T\sigma(\mathbf{b}\mathbf{x} + \mathbf{c})| \le \|\mathbf{a}\|_*(\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|) \text{ for all } (\mathbf{a}, \mathbf{b}, \mathbf{c}) \in \mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V \text{ and } \mathbf{x} \in \Omega. \tag{3.40}$$

Our illustrative examples of the norm $\|\cdot\|$ and its dual norm $\|\cdot\|_*$ include the ℓ^p-norm $\|\cdot\|_p$ and its dual norm $\|\cdot\|_q$, especially the bounded norm $\|\cdot\|_\infty$ and its dual norm $\|\cdot\|_1$ in the distributed implementation and efficient learning of shallow GCNNs, where $1/p + 1/q = 1$.

Remark 3.8 Unlike the classical Barron space as introduced in [3], the Barron space discussed in [2, 17, 19] depends on the choice of the activation function. In this paper, we use the ReLU activation function $\sigma(t) = \max(0, t)$ for the graph Barron space $\mathcal{B}$ for GCNNs. An important property of the ReLU function is its homogeneity:

$$\sigma(\lambda t) = \lambda\sigma(t) \quad \text{for all } \lambda \ge 0 \text{ and } t \in \mathbb{R}.$$

A trivial function satisfying the above homogeneous property is the linear function $\sigma(t) = t$. In this case, the corresponding graph Barron space encompasses all affine functions:

$$f(\mathbf{x}) = \mathbf{v}^T\mathbf{x} + w, \ \mathbf{x} \in \Omega,$$

where $\mathbf{v} \in \mathbb{R}^V$ and $w \in \mathbb{R}$, provided that $\mathcal{W}$ is the entire space $\mathbb{R}^V$. Moreover, one may verify that

$$\|f\|_{\mathcal{B}} \le \|\mathbf{I}_0\|_{\mathrm{co}}\|\mathbf{v}\|_* + |w|, \tag{3.41}$$

where $\mathbf{I}_0 \in \mathbb{R}^V$ is chosen such that the corresponding convolution operator $\mathbf{I}_0 *$ is the identity operator if $\mathbf{v} \neq 0$ and $\mathbf{I}_0 = \mathbf{0}$ if $\mathbf{v} = 0$. Hence the Barron norm on the set of all affine functions defines an equivalent norm for the coefficient vector $[\mathbf{v}^T, w]^T \in \mathbb{R}^V \times \mathbb{R}$ of the affine function f. The estimate in (3.41) holds as we can select the delta measure on some $(\mathbf{a}_0, \mathbf{b}_0, \mathbf{c}_0) \in \mathbb{R}^{3N}$ as the probability measure in the representation (3.5) for the affine function f, where $\mathbf{a}_0 = \mathbf{u}_0$, $\mathbf{b}_0 = \mathbf{I}_0$ and $\mathbf{c}_0 \in \mathbb{R}^V$ is constructed so that $\mathbf{a}_0^T \mathbf{c}_0 = w$ and $\|\mathbf{a}_0\|_* \|\mathbf{c}_0\| = |w|$.

For the ReLU activation function σ, as $t = \sigma(t) + \sigma(-t)$ for all $t \in \mathbb{R}$, the proposed graph Barron space $\mathcal{B}$ with the ReLU as the activation function contains all affine functions on the domain Ω, and

$$\|f\|_{\mathcal{B}} \leq 2\|\mathbf{I}_0\|_{\mathrm{co}} \|\mathbf{v}\|_* + 2|w|, \tag{3.42}$$

for all affine functions $f = \mathbf{v}^T \mathbf{x} + w$ with $\mathbf{v} \in \mathbb{R}^V$ and $w \in \mathbb{R}$.

Remark 3.9 Let Ω be a compact domain of $\mathbb{R}^V$, φ_v, $v \in \mathcal{V}$ be a collection of neurons such that $\sup_{v \in \mathcal{V}} |\varphi_v(\mathbf{x})| < \infty$ for all $\mathbf{x} \in \Omega$, and Υ be a family of probability measures $\hat{\rho}$ on $\mathcal{V}$. For functions f admitting the representation

$$f(\mathbf{x}) = c \int_{v \in \mathcal{V}} \varphi_v(\mathbf{x}) \hat{\rho}(dv), \ \mathbf{x} \in \Omega, \tag{3.43}$$

with $c \in \mathbb{R}$ and $\hat{\rho} \in \Upsilon$, we have the following pointwise estimate:

$$|f(\mathbf{x})| \leq |c| \sup_{v \in \mathcal{V}} |\varphi_v(\mathbf{x})|, \ \mathbf{x} \in \Omega. \tag{3.44}$$

Then taking infimum over all probability measures $\hat{\rho} \in \Upsilon$ leads to the reproducing kernel property:

$$|f(\mathbf{x})| \leq \sup_{v \in \mathcal{V}} |\varphi_v(\mathbf{x})| \|f\|_{\tilde{\mathcal{B}}} \tag{3.45}$$

where

$$\|f\|_{\tilde{\mathcal{B}}} = \inf\{|c| : \ f \text{ satisfies (3.43) for some probability measure } \hat{\rho} \in \Upsilon\}.$$

The reproducing kernel property (3.12) for the graph Barron space $\mathcal{B}$ is derived from the representation (3.8) of functions, combined with the neuron estimate (3.40) and the bound estimate (3.44) for the evaluation functional. It is noteworthy that a similar reproducing kernel property has been established for the Barron space B_{EMW} in the classical neural network setting [17, 19, 63].

3.2 Reproducing Kernel Hilbert Spaces with Neuron Kernels

For $\hat{\rho} \in \widehat{P}$, let $L^p_{\hat{\rho}} := L^p_{\hat{\rho}}(\mathbb{S} \times \mathbb{T})$, $1 \le p < \infty$, be the Banach space of all p-integrable functions on $\mathbb{S} \times \mathbb{T}$ with respect to the probability measure $\hat{\rho}$ and define its norm by

$$\|\eta\|_{L^p_{\hat{\rho}}} = \Big(\int_{\mathbb{S}\times\mathbb{T}} |\eta(\mathbf{a}, \mathbf{b}, \mathbf{c})|^p \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \Big)^{1/p}.$$

Denote the completion of the linear space spanned by $\mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c})$, $\mathbf{x} \in \Omega$ in $L^2_{\hat{\rho}}$ by $\mathcal{L}^2_{\hat{\rho}} := \mathcal{L}^2_{\hat{\rho}}(\mathbb{S} \times \mathbb{T})$, and let $P_{\hat{\rho}}$ denote the orthogonal projection from $L^2_{\hat{\rho}}$ onto its Hilbert subspace $\mathcal{L}^2_{\hat{\rho}}$. In the following theorem, we show that a function f in the RKHS $\mathcal{H}_{\hat{\rho}}$ can be represented by some function $\eta \in L^2_{\hat{\rho}}(\mathbb{S} \times \mathbb{T})$, and we establish both the corresponding norm characterization and the reproducing kernel property.

Theorem 3.10 *Let $\hat{\rho} \in \widehat{P}$ be a probability measure on $\mathbb{S} \times \mathbb{T}$, $P_{\hat{\rho}}$ be the orthogonal projection from $L^2_{\hat{\rho}}$ onto its subspace $\mathcal{L}^2_{\hat{\rho}}$, and $\mathcal{H}_{\hat{\rho}}$ be the RKHS with kernel $K_{\hat{\rho}}$ given in (3.10) and norm denoted by $\|\cdot\|_{\mathcal{H}_{\hat{\rho}}}$. Then $g \in \mathcal{H}_{\hat{\rho}}$ if and only if*

$$g(\mathbf{x}) = \int_{\mathbb{S}\times\mathbb{T}} \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}) \eta(\mathbf{a}, \mathbf{b}, \mathbf{c}) \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \tag{3.46}$$

for some function $\eta \in L^2_{\hat{\rho}}$. Moreover,

$$\|g\|_{\mathcal{H}_{\hat{\rho}}} = \|P_{\hat{\rho}}\eta\|_{L^2_{\hat{\rho}}} \tag{3.47}$$

and

$$\|g\|_\infty \le \|g\|_{\mathcal{H}_{\hat{\rho}}}. \tag{3.48}$$

Proof Take $\eta \in L^2_{\hat{\rho}}$ and let $g(\mathbf{x})$ be as in (3.46). Then the function g is a bounded function on the domain Ω, since

$$\begin{aligned} |g(\mathbf{x})| &\le \int_{\mathbb{S}\times\mathbb{T}} |\mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c})| |\eta(\mathbf{a}, \mathbf{b}, \mathbf{c})| \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \\ &\le \int_{\mathbb{S}\times\mathbb{T}} |\eta(\mathbf{a}, \mathbf{b}, \mathbf{c})| \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \le \|\eta\|_{L^2_{\hat{\rho}}}, \ \mathbf{x} \in \Omega. \end{aligned} \tag{3.49}$$

Denote the inner product on $L^2_{\hat{\rho}}$ by $\langle \cdot, \cdot \rangle_{\hat{\rho}}$, and set $\tilde{\eta} = P_{\hat{\rho}}\eta \in \mathcal{L}^2_{\hat{\rho}}$, the orthogonal projection of the function $\eta \in L^2_{\hat{\rho}}(\mathbb{S} \times \mathbb{T})$ onto $\mathcal{L}^2_{\hat{\rho}}(\mathbb{S} \times \mathbb{T})$. Then for any $\mathbf{x}_0 \in \Omega$,

$$g(\mathbf{x}_0) - \int_{\mathbb{S}\times\mathbb{T}} \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x}_0 + \mathbf{c}) \tilde{\eta}(\mathbf{a}, \mathbf{b}, \mathbf{c}) \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) = \langle \eta - \tilde{\eta}, \psi(\mathbf{x}_0, \cdot) \rangle_{\hat{\rho}} = 0, \tag{3.50}$$

where $\psi\big(\mathbf{x}, (\mathbf{a}, \mathbf{b}, \mathbf{c})\big) = \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c})$ and the last equality follows since $\eta - \tilde{\eta}$ is orthogonal to $\mathcal{L}^2_{\hat{\rho}}$ and $\psi(\mathbf{x}_0, \cdot) \in \mathcal{L}^2_{\hat{\rho}}$ for every $\mathbf{x}_0 \in \Omega$. By (3.50) and

$$\|\eta\|^2_{L^2_{\hat{\rho}}} = \|\tilde{\eta}\|^2_{L^2_{\hat{\rho}}} + \|\eta - \tilde{\eta}\|^2_{L^2_{\hat{\rho}}},$$

it remains to establish (3.46), (3.47), and (3.48) for $\eta \in \mathcal{L}^2_{\hat{\rho}}$.

Let $\mathcal{H}^o_{\hat{\rho}}$ be the linear span of $K_{\hat{\rho}}(\cdot, \mathbf{x}')$, $\mathbf{x}' \in \Omega$, and define the inner product on $\mathcal{H}^o_{\hat{\rho}}$ between $g_1 = \sum_{j=1}^J b_j K_{\hat{\rho}}(\cdot, \mathbf{x}_j) \in \mathcal{H}^o_{\hat{\rho}}$ and $g_2 = \sum_{i=1}^I a_i K_{\hat{\rho}}(\cdot, \mathbf{x}'_i) \in \mathcal{H}^o_{\hat{\rho}}$ by

$$\langle g_1, g_2 \rangle_{\mathcal{H}_{\hat{\rho}}} = \sum_{i=1}^I \sum_{j=1}^J a_i b_j K_{\hat{\rho}}(\mathbf{x}_j, \mathbf{x}'_i). \tag{3.51}$$

Then one may verify that $g \in \mathcal{H}^o_{\hat{\rho}}$ if and only if $g = \sum_{i=1}^I c_i K_{\hat{\rho}}(\cdot, \mathbf{x}_i)$ for some $c_i \in \mathbb{R}$ and $\mathbf{x}_i \in \Omega$, $1 \le i \le I$, if and only if

$$g(\mathbf{x}) = \int_{\mathbb{S}\times\mathbb{T}} \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}) \eta(\mathbf{a}, \mathbf{b}, \mathbf{c}) \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \tag{3.52}$$

for some function $\eta \in \mathcal{L}^o$, the linear space spanned by $\mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c})$, $\mathbf{x} \in \Omega$. Moreover,

$$\|g\|_{\mathcal{H}_{\hat{\rho}}} = \Big(\int_{\mathbb{S}\times\mathbb{T}} |\eta(\mathbf{a}, \mathbf{b}, \mathbf{c})|^2 \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \Big)^{1/2} \tag{3.53}$$

by (3.51), and

$$|g(\mathbf{x})| \le \Big(\int_{\mathbb{S}\times\mathbb{T}} |\eta(\mathbf{a}, \mathbf{b}, \mathbf{c})|^2 \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \Big)^{1/2} = \|g\|_{\mathcal{H}_{\hat{\rho}}}, \ \mathbf{x} \in \Omega. \tag{3.54}$$

This proves (3.46), (3.47) and (3.48) for functions $g \in \mathcal{H}^o_{\hat{\rho}}$. Recall that $\mathcal{H}_{\hat{\rho}}$ and $\mathcal{L}^2_{\hat{\rho}}$ are the completion of $\mathcal{H}^o_{\hat{\rho}}$ and $\mathcal{L}^o$ respectively. Hence taking limits in (3.52), (3.53) and (3.54) proves the desired conclusion (3.46), (3.47) and (3.48) with $\eta \in \mathcal{L}^2_{\hat{\rho}}$, and hence completes the proof. □

Remark 3.11 We remark that representing function $\eta(\mathbf{a}, \mathbf{b}, \mathbf{c}) \in \mathcal{L}^2_{\hat{\rho}}$ for the RKHS $\mathcal{H}_{\hat{\rho}}$ is linear with respect to $\mathbf{a} = [a_v]_{v\in V}$, i.e.,

$$\eta(\mathbf{a}, \mathbf{b}, \mathbf{c}) = \sum_{v\in V} a_v \tau_v(\mathbf{b}, \mathbf{c}) \tag{3.55}$$

for some functions τ_v, $v \in V$, on $\mathbb{T}$. Let $\mathbf{e}_v \in \mathbb{R}^V$, $v \in V$, be the delta graph signal taking value zero all vertices except value one at the vertex v. Observe that

$$|\mathbf{e}_v^T (\mathbf{b} * \mathbf{x})| \le \|\mathbf{e}_v\|_* \|\mathbf{b} * \mathbf{x}\| \le \|\mathbf{e}_v\|_* \|\mathbf{b}\|_{\mathrm{co}}, \ v \in V \text{ and } \mathbf{x} \in \Omega.$$

Therefore in addition to the linearity with respect to $\mathbf{a}$ for the representing function η in the RKHS $\mathcal{H}_{\hat{\rho}}$, the functions τ_v, $v \in V$, in (3.55) are linear with respect to $\mathbf{b}$ and $\mathbf{c}$ in the domain $\{(\mathbf{b}, \mathbf{c}) \in \mathbb{T} : \ \|\mathbf{e}_{v'}\|_* \|\mathbf{b}\|_{\mathrm{co}} \le |\mathbf{e}_v^T \mathbf{c}|, v' \in V\}$.

3.3 Graph Barron Space and RKHSs with Neuron Kernels

In the following theorem, we show that every function in the Barron space $\mathcal{B}$ belongs to some RKHS $\mathcal{H}_{\hat{\rho}}$, $\hat{\rho} \in \widehat{P}$, and conversely, that every function in the RKHS $\mathcal{H}_{\hat{\rho}}$, $\hat{\rho} \in \widehat{P}$, belongs to the Barron space $\mathcal{B}$. Moreover, we establish the equivalence of the corresponding norms.

Theorem 3.12 *Let $\mathcal{B}$ be the graph Barron space in (3.7), and $\mathcal{H}_{\hat{\rho}}$, $\hat{\rho} \in \widehat{\mathcal{P}}$, be RKHSs with kernels given in (3.10). Then*

$$\mathcal{B} = \cup_{\hat{\rho} \in \widehat{P}} \mathcal{H}_{\hat{\rho}} \tag{3.56}$$

and

$$\|f\|_{\mathcal{B}} = \inf_{f \in H_{\hat{\rho}},\ \hat{\rho} \in \widehat{P}} \|f\|_{\mathcal{H}_{\hat{\rho}}}, \ \ f \in \mathcal{B}. \tag{3.57}$$

Proof Take $f \in \mathcal{B}$ and let $\hat{\rho} \in \widehat{P}$ be the probability measure on $\mathbb{S} \times \mathbb{T}$ such that (3.9) holds. Then by Theorem 3.10, we conclude that $f \in \mathcal{H}_{\hat{\rho}}$ and

$$\|f\|_{\mathcal{H}_{\hat{\rho}}} \le \|f\|_{\mathcal{B}} \Big(\int_{\mathbb{S} \times \mathbb{T}} |(P_{\hat{\rho}} 1)(\mathbf{a}, \mathbf{b}, \mathbf{c})|^2 \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \Big)^{1/2} \le \|f\|_{\mathcal{B}}.$$

This shows that

$$\mathcal{B} \subset \cup_{\hat{\rho} \in \widehat{P}} \mathcal{H}_{\hat{\rho}} \ \text{ and } \ \inf_{f \in H_{\hat{\rho}}, \hat{\rho} \in \widehat{P}} \|f\|_{\mathcal{H}_{\hat{\rho}}} \le \|f\|_{\mathcal{B}}. \tag{3.58}$$

Let $f \in \mathcal{H}_{\hat{\rho}}$ for some $\hat{\rho} \in \widehat{P}$ and $\eta \in \mathcal{L}^2_{\hat{\rho}}$ so that (3.46) holds. The existence of such a function η follows from Theorem 3.10. Moreover, we have

$$\|f\|_{\mathcal{H}_{\hat{\rho}}} = \|\eta\|_{L^2_{\hat{\rho}}}. \tag{3.59}$$

Define a probability measure $\hat{\tilde{\rho}}$ on $\mathbb{S} \times \mathbb{T}$ by

$$\hat{\tilde{\rho}}(A) = \frac{\int_{E_1 \cap A} \eta(\mathbf{a}, \mathbf{b}, \mathbf{c}) \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) - \int_{E_2 \cap \tilde{A}} \eta(\mathbf{a}, \mathbf{b}, \mathbf{c}) \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c})}{\|\eta\|_{L^1_{\hat{\rho}}}}, \tag{3.60}$$

where $E_1 = \{(\mathbf{a}, \mathbf{b}, \mathbf{c}) \in \mathbb{S} \times \mathbb{T} : \ \eta(\mathbf{a}, \mathbf{b}, \mathbf{c}) \ge 0\}$, $E_2 = \{(\mathbf{a}, \mathbf{b}, \mathbf{c}) \in \mathbb{S} \times \mathbb{T} : \eta(\mathbf{a}, \mathbf{b}, \mathbf{c}) < 0\}$, and $\tilde{A} = \{(\mathbf{a}, \mathbf{b}, \mathbf{c}) \in \mathbb{S} \times \mathbb{T} : \ (-\mathbf{a}, \mathbf{b}, \mathbf{c}) \in A\}$. By (3.46) and the

definition (3.60) of the probability measure $\hat{\tilde{\rho}} \in \hat{P}$, we have

$$\begin{aligned} f(\mathbf{x}) &= \int_{E_1} \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}) \eta(\mathbf{a}, \mathbf{b}, \mathbf{c}) \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \\ &\quad + \int_{E_2} (-\mathbf{a})^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c})(-\eta)(\mathbf{a}, \mathbf{b}, \mathbf{c}) \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \\ &= \|\eta\|_{L^1_{\hat{\rho}}} \int_{\mathbb{S}\times\mathbb{T}} \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}) \hat{\tilde{\rho}}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}). \end{aligned}$$

This implies that $f \in \mathcal{B}$ and $\|f\|_{\mathcal{B}} \leq \|\eta\|_{L^1_{\hat{\rho}}}$. This together with (3.59) and the observation $\|\eta\|_{L^1_{\hat{\rho}}} \leq \|\eta\|_{L^2_{\hat{\rho}}}$ implies that

$$\mathcal{H}_{\hat{\rho}} \subset \mathcal{B} \text{ and } \|f\|_{\mathcal{B}} \leq \|f\|_{\mathcal{H}_{\hat{\rho}}} \text{ for all } f \in \mathcal{H}_{\hat{\rho}}. \tag{3.61}$$

Combining (3.59) and (3.61) completes the proof. □

Remark 3.13 In the standard neural network setting, a similar conclusion to the one in (3.56) about RKHSs and the Barron space is given in [17, Proposition 3], however the corresponding norm equivalence in (3.57) is not mentioned.

4 Approximation Theorems on Graph Barron Spaces

Let $\mathcal{G} = (V, E)$ be a undirected graph of order N. Given a shallow GCNN with parameter $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M) \in (\mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V)^M$, we define its *p-path norm* by

$$\|\boldsymbol{\Theta}\|_{P,p} = \begin{cases} \left(M^{-1} \sum_{m=1}^M \|\boldsymbol{\theta}_m\|^p\right)^{1/p} & \text{if } 1 \leq p < \infty \\ \sup_{1\leq m\leq M} \|\boldsymbol{\theta}_m\| & \text{if } p = \infty, \end{cases} \tag{4.1}$$

where $\|\boldsymbol{\theta}\| = \|\mathbf{a}\|_*(\|\mathbf{b}\|_{\mathrm{co}} + \|\mathbf{c}\|)$ for $\boldsymbol{\theta} = (\mathbf{a}, \mathbf{b}, \mathbf{c}) \in \mathbb{R}^V \times \mathcal{W} \times \mathbb{R}^V$, cf. [17] for $p = 1$. One may verify that the output of the shallow GCNN with parameter $\boldsymbol{\Theta}$ belongs to the Barron space $\mathcal{B}$,

$$\left\| \frac{1}{M} \sum_{m=1}^M \phi(\mathbf{x}, \boldsymbol{\theta}_m) \right\|_{\mathcal{B}} \leq \frac{1}{M} \sum_{m=1}^M \|\boldsymbol{\theta}_m\| \leq \|\boldsymbol{\Theta}\|_{P,p}, 1 \leq p \leq \infty, \tag{4.2}$$

where $\phi(\mathbf{x}, \boldsymbol{\theta}_m) = \mathbf{a}_m^T \sigma(\mathbf{b}_m * \mathbf{x} + \mathbf{c}_m)$ for $\boldsymbol{\theta}_m = (\mathbf{a}_m, \mathbf{b}_m, \mathbf{c}_m), 1 \leq m \leq M$. In the classical neural network setting, functions in Barron/Besov/Hölder spaces can be well approximated by outputs of some neural networks, see [17, 58, 61, 75, 76] and references therein. In Section 4.1, we show that functions in the graph Barron space $\mathcal{B}$ can be approximated in the integrated square norm by suitably chosen shallow GCNNs with limited number of neurons per vertex. In particular, a shallow GCNN can be selected so that its path norm is bounded by the Barron norm of the target function, and the approximation error bound scales linearly with both the square root

of the number M of neurons per vertex and the Barron norm of the target function; see Theorem 4.1. As a consequence, for any $f \in \mathcal{B}$ on a undirected graph of size N, an integrated square error $\epsilon > 0$ can be achieved by a shallow GCNN whose total number of parameters satisfies $3MN > 3N\epsilon^{-2}\|f\|_{\mathcal{B}}^2$ and whose path norm is bounded by $\|f\|_{\mathcal{B}}$.

For any $\epsilon > 0$, we say that the family of balls $B(\mathbf{x}_i, \epsilon)$ with radius ϵ and center $\mathbf{x}_i \in \Omega$, $1 \le i \le I$, is a ϵ-*covering* of the domain Ω if

$$\Omega \subset \cup_{i=1}^{I} B(\mathbf{x}_i, \epsilon), \tag{4.3}$$

and define the ϵ-*covering number* N_ϵ^{ext} by the minimal number of balls in a ϵ-covering of the domain Ω. In Section 4.2, based on the covering of the domain Ω with balls of small radius and the Lipschitz property for functions in the Barron space, we establish a uniform approximation theorem for functions in the Barron space on the domain Ω by outputs of some shallow GCNNs with bounded path norm; see Theorem 4.3. As a consequence of Theorem 4.3, regarding approximation properties via shallow GCNNs, functions in the graph Barron space exhibit significantly different behavior on the conventional domain Ω_{ba} of all bounded signals in (2.16) and the domain Ω_{sp} of all s-sparse signals in (2.15). In particular, for any function f in the graph Barron space on the s-sparse domain Ω_{sp} in (2.15), we can find a shallow GCNN with parameter size $O(sN\epsilon^{-2}\ln(Ns^{-1}\epsilon^{-1}))$ to approximate f uniformly on the domain Ω_{sp} within an accuracy of $\|f\|_{\mathcal{B}}\epsilon$; see Remark 4.4. Applying Theorem 4.3 to the familiar unit ball Ω_{ba} in (2.16), we observe that a shallow GCNN with parameter size $O(N^2\epsilon^{-2}\ln(1/\epsilon))$ can be chosen to approximate any function f in the Barron space uniformly on Ω_{ba} to within an accuracy of $|f|_{\mathcal{B}}\epsilon$; see Remark 4.5 and the detailed comparison with the uniform approximation of functions in a Barron space in the classical neural network setting

For $Q \ge 0$ and $1 \le p \le \infty$, let $\mathcal{F}_Q$ and $\mathcal{C}_{Q,p}$ be the sets of functions in the Barron space with their Barron norms bounded by Q and outputs of all shallow GCNNs with p-path norms of their parameters bounded by Q respectively, i.e.,

$$\mathcal{F}_Q = \{f \in \mathcal{B} : \|f\|_{\mathcal{B}} \le Q\}, \tag{4.4}$$

and

$$\mathcal{C}_{Q,p} = \cup_{M=1}^{\infty} \mathcal{C}_{Q,p,M}, \tag{4.5}$$

where

$$\mathcal{C}_{Q,p,M} = \left\{ \frac{1}{M} \sum_{m=1}^{M} \phi(\mathbf{x}, \boldsymbol{\theta}_m) : \left\| (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M) \right\|_{P,p} \le Q \right\}, \ M \ge 1. \tag{4.6}$$

By (4.2), we see that any function $g \in \mathcal{C}_{Q,p}$ belongs to the Barron space $\mathcal{B}$ and has its Barron norm bounded by Q, i.e.,

$$\|g\|_{\mathcal{B}} \le Q \ \text{ for all } g \in \mathcal{C}_{Q,p}, \tag{4.7}$$

and hence

$$\mathcal{C}_{Q,p} \subset \mathcal{F}_Q. \tag{4.8}$$

As a conclusion of Theorem 4.3, the set $\mathcal{C}_{Q,p}$ has the following density property:

$$\inf_{g \in \mathcal{C}_{Q,p}} \|g(\mathbf{x}) - f(\mathbf{x})\|_\infty = 0 \tag{4.9}$$

hold for all $f \in \mathcal{F}_Q$ and $1 \le p \le \infty$. In Section 4.3, we establish an inverse uniform approximation property for functions in $\mathcal{C}_{Q,p}$, $1 \le p \le \infty$. In particular, we show in Theorem 4.6 that $f \in \mathcal{F}_Q$ if it is the uniform limit of $f_n \in \mathcal{C}_{Q,p}, n \ge 1$; cf. [17, Theorem 2] for a related result in the context of classical neural networks.

Universal approximation theorem is one of fundamental problems in theoretical learning research [7, 31, 32]. In Section 4.4, we establish a universal approximation theorem in the GCNN setting. In particular, we show that the Barron space $\mathcal{B}(\Omega, \mathbb{R}^V)$ is dense in $C(\Omega)$, the space of continuous functions on Ω, provided that all entries of the Fourier transform of some delta graph signal are nonzero; see Theorem 4.7 and cf. Theorem 3.1. Combining the conclusions in Theorems 4.3 and 4.7, we conclude that any continuous function can be **well** approximated by a shallow GCNN when all graph signals are employed in the convolution operation.

Let $\mathbf{U}$ be the orthogonal matrix in (2.3) that simultaneously diagonalizes graph shifts $\mathbf{S}_1, \ldots, \mathbf{S}_d$ and is also used in the definition (2.5) of the graph Fourier transform. Then we can reformulate the nonzero requirement in Theorem 4.7 as the existence of some row of the orthogonal matrix $\mathbf{U}$ with all its components taking nonzero values. We remark that in Theorem 4.7, the above nonzero assumption can **not** be removed. For instance, consider the underlying graph $\mathcal{G} = (V, E)$ being disconnected and select its graph Laplacian as the graph shift to define the convolution operation in our GCNNs. In this scenario, the graph shift can be written as a block diagonal matrix, and consequently the orthogonal matrix that diagonalizes it is also block diagonal and fails to satisfy the non-zero entry requirement in Theorem 4.7. In particular, we can decompose the graph $\mathcal{G}$ into two disconnected subgraphs $\mathcal{G}_1 = (V_1, E_1)$ and $\mathcal{G}_2 = (V_2, E_2)$. Then one can verify that any function f in the Barron space $\mathcal{B}(\Omega, \mathbb{R}^V)$ can be decomposed as the summation of two functions f_1 and f_2 defined on those two disconnected components $\mathcal{G}_1$ and $\mathcal{G}_2$ respectively, i.e.,

$$f(\mathbf{x}) = f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2),$$

where $\mathbf{x} = [x_v]_{v \in V}$ with $\mathbf{x}_1 = [x_v]_{v \in V_1}$ and $\mathbf{x}_2 = [x_v]_{v \in V_2}$ respectively. This separability implies that such functions cannot approximate all continuous functions on the domain Ω and hence the Barron space is not dense in $C(\Omega)$. We remark that the density of the Barron space $\mathcal{B}(\Omega, \mathbb{R}^V)$ in $C(\Omega)$ may not hold in general if the space $\mathcal{W}$ used for convolutions in GCNNs is not the whole space $\mathbb{R}^V$; see Remark 4.9.

4.1 Approximation of GCNNs in Integrated Square Norm

In the following theorem, we show that shallow GCNNs can effectively approximate functions in the graph Barron space $\mathcal{B}$ with respect to the integrated square norm.

Theorem 4.1 *Let $M \geq 1$, $f \in \mathcal{B}$ and μ be a probability measure on the domain Ω. Then for any $\delta > 0$ there is a shallow GCNN with parameter $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$ such that*

$$\|\boldsymbol{\Theta}\|_{P,\infty} \leq \|f\|_{\mathcal{B}} \tag{4.10}$$

and

$$\int_{\Omega} \Big| \frac{1}{M} \sum_{m=1}^{M} \phi(\mathbf{x}, \boldsymbol{\theta}_m) - f(\mathbf{x}) \Big|^2 \mu(d\mathbf{x}) \leq \frac{1+\delta}{M} \|f\|_{\mathcal{B}}^2. \tag{4.11}$$

Proof Without loss of generality, we assume that $f \in \mathcal{B}$ is a nonzero function with $\|f\|_{\mathcal{B}} = 1$, otherwise replacing f by $f/\|f\|_{\mathcal{B}}$. Then by Lemma 3.3, there exists a probability measure $\hat{\rho}$ on $\mathbb{S} \times \mathbb{T}$ such that

$$f(\mathbf{x}) = \int_{\mathbb{S}\times\mathbb{T}} \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}) \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}), \ \ \mathbf{x} \in \Omega. \tag{4.12}$$

Let $\boldsymbol{\theta}_m = (\mathbf{a}_m, \mathbf{b}_m, \mathbf{c}_m) \in \mathbb{S} \times \mathbb{T}$, $1 \leq m \leq M$, be i.i.d. random variables following the probability measure $\hat{\rho}$. Set $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$ and define

$$f_M(\mathbf{x}, \boldsymbol{\Theta}) = \frac{1}{M} \sum_{m=1}^{M} \phi(\mathbf{x}, \boldsymbol{\theta}_m), \ \ \mathbf{x} \in \Omega.$$

Clearly, we have

$$\|\boldsymbol{\Theta}\|_{P,\infty} = 1 = \|f\|_{\mathcal{B}}.$$

Observe that $\phi(\mathbf{x}, \boldsymbol{\theta}_m) = \mathbf{a}_m^T \sigma(\mathbf{b}_m * \mathbf{x} + \mathbf{c}_m)$ are i.i.d random variables with $\phi(\mathbf{x}, \boldsymbol{\theta}_m) \in [-1, 1]$ almost surely and $\mathbb{E}\phi(\mathbf{x}, \boldsymbol{\theta}_m) = f(\mathbf{x})$ hold for all $1 \leq m \leq M$ and $\mathbf{x} \in \Omega$. Combining the above observation with (1.1) and (4.12), we obtain

$$\begin{aligned} \mathbb{E} \int_{\Omega} (f_M(\mathbf{x}, \boldsymbol{\Theta}) - f(\mathbf{x}))^2 d\mu(\mathbf{x}) &= \frac{1}{M} \int_{\Omega} \mathbb{E}(\phi(\mathbf{x}, \boldsymbol{\theta}) - \mathbb{E}\phi(\mathbf{x}, \boldsymbol{\theta}))^2 d\mu(\mathbf{x}) \\ &\leq \frac{1}{M} \int_{\Omega} \mathbb{E}(\phi(\mathbf{x}, \boldsymbol{\theta}))^2 d\mu(\mathbf{x}) \leq \frac{1}{M}. \end{aligned}$$

Applying the Markov's inequality yields

$$\mathbb{P}\Big\{ \int_{\Omega} (f_M(\mathbf{x}, \boldsymbol{\Theta}) - f(\mathbf{x}))^2 d\mu(\mathbf{x}) > \frac{1+\delta}{M} \Big\} \leq \frac{1}{1+\delta} < 1.$$

This completes the proof. □

Taking Dirac measure at some $\mathbf{x}_0 \in \Omega$ as the probability measure in Theorem 4.1, we have the following pointwise estimate.

Corollary 4.2 *Let $M \geq 1$ and $f \in \mathcal{B}$. Then for any $\epsilon > 0$ and $\mathbf{x}_0 \in \Omega$, there is a shallow GCNN with parameter $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$ such that $\|\boldsymbol{\Theta}\|_{P,\infty} \leq \|f\|_{\mathcal{B}}$ and*

$$\left| \frac{1}{M} \sum_{m=1}^{M} \phi(\mathbf{x}_0, \boldsymbol{\theta}_m) - f(\mathbf{x}_0) \right| \leq \frac{1+\epsilon}{\sqrt{M}} \|f\|_{\mathcal{B}}. \tag{4.13}$$

We remark that the shallow GCNN chosen in Corollary 4.2 may depend on $\mathbf{x}_0 \in \Omega$.

4.2 Approximation of GCNNs in Uniform Norm

In the following theorem, we consider the uniform approximation of functions in the Barron space on the domain Ω by shallow GCNNs, provided the number M of neurons at each vertex is sufficiently large.

Theorem 4.3 *Let $\epsilon \in (0, 1/2)$. Assume that the ReLU function σ satisfies (2.17) and (2.18), the convolution norm satisfies (2.19) and (2.20), and the domain Ω has a ϵ-covering in (4.3) with ϵ-covering number N_ϵ^{ext}. If the number $M \geq 1$ of neurons per vertex is chosen to satisfy*

$$2N_\epsilon^{\text{ext}} e^{-M\epsilon^2/2} < 1, \tag{4.14}$$

then for any function f in the Barron space $\mathcal{B}$ there exists a shallow GCNN with parameter $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$ such that

$$\|\boldsymbol{\Theta}\|_{P,\infty} \leq \|f\|_{\mathcal{B}} \tag{4.15}$$

and

$$\left\| \frac{1}{M} \sum_{m=1}^{M} \phi(\cdot, \boldsymbol{\theta}_m) - f \right\|_\infty \leq (2D_1 \|\sigma\|_{\text{Lip}} + 1)\epsilon \|f\|_{\mathcal{B}}, \tag{4.16}$$

where $\|\sigma\|_{\text{Lip}}$ and D_1 are the constants in (2.18) and (2.20) respectively.

Proof of Theorem 4.3 We follow the argument of Theorem 4.1. Without loss of generality, we assume that $\|f\|_{\mathcal{B}} = 1$. Let $\hat{\rho}$ be the probability measure on $\mathbb{S} \times \mathbb{T}$ in (4.12), $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$ be the i.i.d random variables following the probability measure $\hat{\rho}$, and set $f_M(\mathbf{x}, \boldsymbol{\Theta}) = M^{-1} \sum_{m=1}^{M} \phi(\cdot, \boldsymbol{\theta}_m)$.

Set $I = N_\epsilon^{\text{ext}}$ and take a family of balls $B(\mathbf{x}_i, \epsilon)$, $1 \leq i \leq I$, with center $\mathbf{x}_i \in \Omega$ and radius ϵ that covers the domain Ω,

$$\cup_{i=1}^{I} B(\mathbf{x}_i, \epsilon) = \Omega. \tag{4.17}$$

For any random vector $\boldsymbol{\theta} = (\mathbf{a}, \mathbf{b}, \mathbf{c}) \in \mathbb{S} \times \mathbb{T}$, we obtain from (1.1) and the definition of $\mathbb{S} \times \mathbb{T}$ that $\phi(\mathbf{x}_i, \boldsymbol{\theta}) = \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x}_i + \mathbf{c}) \in [-1, 1]$ almost surely and $\mathbb{E}\phi(\mathbf{x}_i, \boldsymbol{\theta}) = f(\mathbf{x}_i)$. Therefore $\phi(\mathbf{x}_i, \theta) - f(\mathbf{x}_i)$ is subGaussian with variance proxy 1,

$$\mathbb{E} \exp\left(s(\phi(\mathbf{x}_i, \boldsymbol{\theta}) - f(\mathbf{x}_i))\right) \leq \exp(s^2/2), \ s \in \mathbb{R}.$$

Therefore for all $t > 0$, we have the following Hoeffding's inequality,

$$\mathbb{P}\{|f_M(\mathbf{x}_i, \boldsymbol{\Theta}) - f(\mathbf{x}_i)| > t\} \le 2\exp(-Mt^2/2). \tag{4.18}$$

This together with (4.14) implies the existence of $\boldsymbol{\Theta}^* \in (\mathbb{S} \times \mathbb{T})^M$ such that

$$\|\boldsymbol{\Theta}^*\|_{P,\infty} \le 1 = \|f\|_{\mathcal{B}} \tag{4.19}$$

and

$$|f_M(\mathbf{x}_i, \boldsymbol{\Theta}^*) - f(\mathbf{x}_i)| \le \epsilon, \ 1 \le i \le I. \tag{4.20}$$

For the above shallow GCNN with parameter $\boldsymbol{\Theta}^*$, we obtain from (4.2), (4.17), (4.20), and Corollary 3.4 that

$$\begin{aligned}
&|f_M(\mathbf{x}, \boldsymbol{\Theta}^*) - f(\mathbf{x})| \\
&\quad \le |f_M(\mathbf{x}, \boldsymbol{\Theta}^*) - f_M(\mathbf{x}_i, \boldsymbol{\Theta}^*)| + |f_M(\mathbf{x}_i, \boldsymbol{\Theta}^*) - f(\mathbf{x}_i)| + |f(\mathbf{x}) - f(\mathbf{x}_i)| \\
&\quad \le D_1\|\sigma\|_{\mathrm{Lip}}\epsilon\|\boldsymbol{\Theta}^*\|_{P,\infty} + \epsilon + D_1\|\sigma\|_{\mathrm{Lip}}\epsilon\|f\|_{\mathcal{B}} \le (2D_1\|\sigma\|_{\mathrm{Lip}} + 1)\epsilon, \ \mathbf{x} \in \Omega,
\end{aligned}$$

where $\mathbf{x}_i$ is chosen so that $\mathbf{x} \in B(\mathbf{x}_i, \epsilon)$. This together with (4.19) completes the proof. □

Remark 4.4 For our illustrative domain Ω_{sp} in (2.15), the ϵ-covering number $N_\epsilon^{\mathrm{ext}}$ is bounded above by $\binom{N}{s}(3/\epsilon)^s$ [66]. This implies that the requirement (4.14) is met if

$$M \ge \frac{2\ln 2}{\epsilon^2} + \frac{2s}{\epsilon^2}\ln\Big(\frac{eN}{s\epsilon}\Big). \tag{4.21}$$

Hence the shallow GCNN is nearly scalable, up to a logarithmic factor in the order N of the underlying graph $\mathcal{G}$, and a shallow GCNN with parameter size $O(sN\epsilon^{-2}\ln(Ns^{-1}\epsilon^{-1}))$ could be chosen to approximate a function f in the graph Barron space uniformly on the domain Ω_{sp} of s-sparse signals with accuracy $\epsilon\|f\|_{\mathcal{B}}$.

Given the number M of neurons per vertex, which is assumed to be much larger than the sparsity parameter s, one can verify that condition (4.21) is satisfied by choosing $\epsilon = (sM^{-1}\ln(NMs^{-2}))^{1/2}$. Consequently, for a function f in the graph Barron space $\mathcal{B}$, the approximation error of shallow GCNNs with M neurons per vertex is of order $\|f\|_{\mathcal{B}}(sM^{-1}\ln(NMs^{-2}))^{1/2}$.

Remark 4.5 For the case that the unit ball Ω_{ba} in (2.16) is used as the domain Ω, we have the following estimate on the ϵ-covering number $(1/\epsilon)^N \le N_\epsilon^{\mathrm{ext}} \le (1 + 2/\epsilon)^N$. Hence the requirement (4.14) is met if

$$M \ge \frac{2\ln 2}{\epsilon^2} + \frac{2N}{\epsilon^2}\ln\Big(1 + \frac{2}{\epsilon}\Big). \tag{4.22}$$

Hence for any $\epsilon \in (0, 1/2)$, shallow GCNNs with parameter size $O(N^2\epsilon^{-2}\ln(1/\epsilon))$ could be selected by Theorem 4.3 to approximate a function f in the Barron space uniformly on the whole domain Ω_{ba} with accuracy $\epsilon\|f\|_{\mathcal{B}}$. However, in the above

setting on the domain, a large number of neurons is necessary for the uniform approximation by shallow GCNNs, it is more appropriate to use multi-layer GCNNs, with each layer containing a smaller number of neurons. In the classical neuron network setting, it is shown in [2, Proposition 1] that functions f in the Barron space $\mathcal{B}_{\mathrm{EMW}}$ in (3.37) can be approximated uniformly on the unit cube $\Omega = [0,1]^N$ by some shadow neural networks with accuracy $\epsilon \|f\|_{\mathcal{B}_{\mathrm{EMW}}}$, provided that

$$M \geq C(N)\epsilon^{-2N/(N+3)} \tag{4.23}$$

where $C(N)$ is the approximation constant of a zonoid by zonotopes with small numbers of segments. The requirements (4.22) and (4.23) on the number of neurons are essentially not comparable because we lack sufficient information about the approximation constant $C(N)$. Furthermore, the order improvement from 2 in (4.22) to $2N/(N+3)$ in (4.23) is not significant, as in designing shallow GCNNs on graphs of large order N, the approximation error ϵ is not chosen making $\epsilon^{-6/(N+3)}$ or equivalent $N^{-1}\ln(1/\epsilon)$ a large number in general.

For $N \geq 2$ and $M \geq 5N$, one may verify that $\epsilon = \sqrt{2(M/N)^{-1}\ln(M/N)}$ satisfies $2(1+2/\epsilon)^N \exp(-M\epsilon^2/2) < 1$. Therefore for any function f in the Barron space $\mathcal{B}$ there exists a shallow GCNN with parameter $\mathbf{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$ such that (4.15) holds and

$$\sup_{\mathbf{x}\in\Omega} \left| \frac{1}{M} \sum_{m=1}^{M} \phi(\mathbf{x}, \boldsymbol{\theta}_m) - f(\mathbf{x}) \right| \leq (2D_1\|\sigma\|_{\mathrm{Lip}} + 1)\sqrt{\frac{2\ln(M/N)}{M/N}}\|f\|_{\mathcal{B}}, \tag{4.24}$$

In the classical neuron network setting with $\Omega = [0,1]^N$, it is mentioned in [18, Remark 13] that there exists a shallow neural network with parameters $u_m \in \mathbb{R}$ and $\mathbf{v}_m \in \mathbb{R}^N$, $1 \leq m \leq M$, such that

$$\sup_{\mathbf{x}\in[0,1]^N} \left| \frac{1}{M} \sum_{m=1}^{M} u_m \sigma(\mathbf{v}_m^T \mathbf{x}) - f(\mathbf{x}) \right| \leq C\sqrt{\frac{\ln M}{M^{1+1/N}}}\|f\|_{\mathcal{B}_{\mathrm{EMW}}}, \tag{4.25}$$

where C is a positive constant depending on N. The approximation error estimates in (4.24) for the GCNN setting and (4.25) for the classical neural networking setting are not comparable due to the lack of information about the constant C. In addition, the order improvement $M^{-1/(2N)}$ in (4.25) is not indicative, as when we design shallow GCNNs on graphs of large order N, the number of neurons M at each vertex is generally chosen so that it does not exceed a multiple of the graph order N (and hence $N^{-1}\ln M$ is not a big number).

4.3 Inverse Uniform Approximation Theorems

In the following theorem, we show that the uniform limit of outputs of shallow GCNNs with bounded path norm belongs to the Barron space $\mathcal{B}$; see [17, Theorem 2] for a similar result in the classical neural network setting.

Theorem 4.6 *Let $Q > 0$ and $1 \le p \le \infty$. If $f_n \in \mathcal{C}_{Q,p}$, $n \ge 1$, converges pointwise, i.e.,*

$$\lim_{n\to\infty} f_n(\mathbf{x}) = f(\mathbf{x}),\ \mathbf{x} \in \Omega, \tag{4.26}$$

for some function f on the domain Ω. Then $f \in \mathcal{B}$ and $\|f\|_{\mathcal{B}} \le Q$.

Proof By the assumption $f_n \in \mathcal{C}_{Q,p}$, $n \ge 1$, we can write

$$f_n(x) = \frac{1}{M_n} \sum_{m=1}^{M_n} \alpha_{n,m} \mathbf{a}_{n,m}^T \sigma(\mathbf{b}_{n,m} * \mathbf{x} + \mathbf{c}_{n,m}),\ \mathbf{x} \in \Omega, \tag{4.27}$$

where $(\mathbf{a}_{n,m}, \mathbf{b}_{n,m}, \mathbf{c}_{n,m}) \in \mathbb{S} \times \mathbb{T}$, $1 \le m \le M_n$, and $0 < M_n^{-1} \sum_{m=1}^{M_n} \alpha_{n,m} \le Q$. Define the probability measure $\hat{\rho}_n$ on $\mathbb{S} \times \mathbb{T}$ by

$$\hat{\rho}_n(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) = \Big(\sum_{m=1}^{M_n} \alpha_{n,m} \Big)^{-1} \sum_{m=1}^{M_n} \alpha_{n,m} \delta\big(\mathbf{a} - \mathbf{a}_{m,n}, \mathbf{b} - \mathbf{b}_{m,n}, \mathbf{c} - \mathbf{c}_{m,n}\big),$$

where δ is the Dirac measure centered at the origin. Then we rewrite the representation in (4.27) as follows:

$$f_n(\mathbf{x}) = \frac{\sum_{m=1}^{M_n} \alpha_{n,m}}{M_n} \int_{\mathbb{S}\times\mathbb{T}} \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}) \hat{\rho}_n(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}),\ \mathbf{x} \in \Omega. \tag{4.28}$$

As $\sum_{m=1}^{M_n} \alpha_{n,m}/M_n$, $n \ge 1$, is a bounded sequence contained in $[0, Q]$, without loss of generality, we assume that it is convergent,

$$\lim_{n\to\infty} \frac{1}{M_n} \sum_{m=1}^{M_n} \alpha_{n,m} = \alpha \in [0, Q], \tag{4.29}$$

otherwise replacing f_n, $n \ge 1$, by its appropriate subsequence.

Consider the sequence $\hat{\rho}_n$, $n \ge 1$, of probability measures on $\mathbb{S} \times \mathbb{T}$. By applying Prokhorov's theorem [50], there exists a subsequence of probability measures $\hat{\rho}_n$ on $\mathbb{S} \times \mathbb{T}$ that converges weakly to some measure $\hat{\rho}$ on $\mathbb{S} \times \mathbb{T}$. Without loss of generality, we assume that the original sequence $\hat{\rho}_n$, $n \ge 1$, of probability measures converges weakly to some measure $\hat{\rho}$ on $\mathbb{S} \times \mathbb{T}$, i.e.,

$$\lim_{n\to\infty} \int_{\mathbb{S}\times\mathbb{T}} g(\mathbf{a}, \mathbf{b}, \mathbf{c}) \hat{\rho}_n(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) = \int_{\mathbb{S}\times\mathbb{T}} g(\mathbf{a}, \mathbf{b}, \mathbf{c}) \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \tag{4.30}$$

hold for all continuous functions g on $\mathbb{S} \times \mathbb{T}$.

As $\hat{\rho}_n$, $n \ge 1$, are probability measures on $\mathbb{S} \times \mathbb{T}$ and the constant function is a continuous function on $\mathbb{S} \times \mathbb{T}$, we obtain from (4.30) that $\hat{\rho}$ is also a probability

measure on $\mathbb{S} \times \mathbb{T}$. Define

$$\tilde{f}(\mathbf{x}) = \alpha \int_{\mathbb{S}\times\mathbb{T}} \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}) \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}), \ \mathbf{x} \in \Omega. \tag{4.31}$$

Then the function $\tilde{f}$ defined by (4.31) belongs to the Barron space $\mathcal{B}$ and has its norm bounded by $\alpha \leq Q$.

Recall that $\mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c})$ are continuous functions about $(\mathbf{a}, \mathbf{b}, \mathbf{c}) \in \mathbb{S} \times \mathbb{T}$ for all $\mathbf{x} \in \Omega$. This together with (4.28), (4.29), (4.30) and (4.31) implies that

$$\tilde{f}(\mathbf{x}) = \lim_{n\to\infty} f_n(\mathbf{x}), \ \mathbf{x} \in \Omega. \tag{4.32}$$

This proves that $\tilde{f} = f$ and hence completes the proof. □

4.4 Universal Approximation Theorem

Given any $v \in V$, denote the delta graph signal taking value one at the vertex v and value zero at all other vertices in V by $\mathbf{e}_v$. In the following theorem, we establish a universal approximation theorem for the Barron space $\mathcal{B}(\Omega, \mathbb{R}^V)$.

Theorem 4.7 *If there exists $v_0 \in V$ such that the Fourier transform $\mathcal{F}\mathbf{e}_{v_0}$ of the delta signal $\mathbf{e}_{v_0}$ has all entries taking nonzero values, then the Barron space $\mathcal{B}(\Omega, \mathbb{R}^V)$ is dense in $C(\Omega)$.*

To prove Theorem 4.7, we recall the following classical universal approximation theorem for neural networks, adapting it from its original formulation on a bounded subset of vectors in $\mathbb{R}^N$ to the bounded domain Ω of graph signals on the graph $\mathcal{G}$ [49].

Lemma 4.8 *Let f be a continuous function on the domain Ω. Then for any $\epsilon > 0$, there exist $u_m \in \mathbb{R}$, $\mathbf{v}_m \in \mathbb{R}^V$ and $w_m \in \mathbb{R}$, $1 \leq m \leq M$ such that*

$$\left\| f(\mathbf{x}) - \sum_{m=1}^{M} u_m \sigma(\mathbf{v}_m^T \mathbf{x} + w_m) \right\|_\infty \leq \epsilon. \tag{4.33}$$

Now we prove Theorem 4.7.

Proof of Theorem 4.7 By the universal approximation theorem in Lemma 4.8, it suffices to prove that for any $\mathbf{v} \in \mathbb{R}^V$ and $w \in \mathbb{R}$ there exist $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^V$ such that

$$\sigma(\mathbf{v}^T \mathbf{x} + w) = \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}), \ \mathbf{x} \in \Omega. \tag{4.34}$$

Take $\mathbf{a} = \mathbf{e}_{v_0}$ and $\mathbf{c} = w\mathbf{e}_{v_0}$. Then the existence of the vector $\mathbf{b}$ in (4.34) is reduced to showing

$$\mathbf{e}_{v_0}^T (\mathbf{b} * \mathbf{x}) = \mathbf{v}^T \mathbf{x} \ \text{ for all } \mathbf{x} \in \mathbb{R}^V. \tag{4.35}$$

By (2.10), we need to find a multivariate polynomial h such that

$$h(\mathbf{S}_1, \ldots, \mathbf{S}_K)\mathbf{e}_{v_0} = \mathbf{v},$$

or equivalently,

$$\operatorname{diag}\big(h(\boldsymbol{\lambda}(1)), \ldots, h(\boldsymbol{\lambda}(N))\big)\mathcal{F}\mathbf{e}_{v_0} = \mathcal{F}\mathbf{v} \tag{4.36}$$

in the Fourier domain. The above equation about the polynomial h is solvable as all entries of $\mathcal{F}\mathbf{e}_{v_0}$ are nonzero and $\boldsymbol{\lambda}(n)$, $1 \le n \le N$, in the joint spectrum of graph shifts are distinct by Assumption 2.1. In particular, h is an interpolation polynomial satisfying

$$h(\boldsymbol{\lambda}(n)) = \frac{\big(\mathcal{F}\mathbf{v}\big)(n)}{\big(\mathcal{F}\mathbf{e}_{v_0}\big)(n)}, \ 1 \le n \le N$$

where $\mathbf{y}(n)$ is the n-th component of a vector $\mathbf{y} \in \mathbb{R}^N$. This completes the proof. □

In the following remark, we consider the density of the graph Barron space $\mathcal{B}(\Omega, \mathcal{W})$ in the space $C(\Omega)$ when $\mathcal{W}$ is not the whole space $\mathbf{R}^V$.

Remark 4.9 Take a nonnegative integer L strictly less than the diameter of the graph $\mathcal{G}$, let $\mathcal{W}_L$ be the space of all polynomial filters with degree no larger than L, and consider replacing the whole space $\mathbb{R}^V$ in Theorem 4.7 by the set $\mathcal{W}_L$ in which the convolutions in our shallow GCNNs are selected.

Take two vertices v_1 and $v_2 \in V$ such that their distance strictly larger than L and define the continuous function g on the domain Ω_{ba} by $g(\mathbf{x}) = x(v_1)x(v_2)$, where $\mathbf{x} = (x(v))_{v\in V}$. Take an arbitrary small $\epsilon > 0$ and suppose that the function g is well approximated by functions in the graph Barron space $f_\epsilon \in \mathcal{B}(\Omega_{\text{ba}}, \mathcal{W}_L)$, i.e.,

$$\sup_{\mathbf{x}\in\Omega_{\text{ba}}} |g(\mathbf{x}) - f_\epsilon(\mathbf{x})| \le \epsilon. \tag{4.37}$$

By (3.9), one may verify that the function f_ϵ can be decomposed as the summation of two functions f_1 and f_2 on $V\backslash\{v_1\}$ and $V\backslash\{v_2\}$ respectively, i.e.,

$$f_\epsilon(\mathbf{x}) = f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) \tag{4.38}$$

where $\mathbf{x}_1 = (x(v))_{v\in V\backslash\{v_1\}}$ and $\mathbf{x}_2 = (x(v))_{v\in V\backslash\{v_2\}}$. Set $\mathbf{x}_1^0 = \mathbf{x}_1$ with $x(v_2) = 0$ and $\mathbf{x}_2^0 = \mathbf{x}_2$ with $x(v_1) = 0$. Then it follows from (4.37) and (4.38) that

$$|f_1(\mathbf{x}_1^0) + f_2(\mathbf{x}_2)| \le \epsilon \ \text{ and } \ |f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2^0)| \le \epsilon. \tag{4.39}$$

Combining (4.37), (4.38) and (4.39) yields

$$|g(\mathbf{x}) - f_1(\mathbf{x}_1^0) - f_2(\mathbf{x}_2^0)| \le 3\epsilon, \ \ \mathbf{x} \in \Omega_{\text{ba}}.$$

This is a contradiction, since the function $f_1(\mathbf{x}_1^0) + f_2(\mathbf{x}_2^0)$ is independent on $x(v_1)$ and $x(v_2)$ and $\epsilon > 0$ is arbitrary chosen. Therefore, unlike the density result for the graph Barron space $\mathcal{B}(\Omega, \mathbb{R}^V)$ in $C(\Omega)$, we conclude that the graph Barron space $\mathcal{B}(\Omega_{\text{ba}}, \mathcal{W}_L)$ is not dense in $C(\Omega_{\text{ba}})$.

5 Rademacher Complexity

Rademacher complexity measures richness of a function class and it has been used to derive data-dependent upper-bounds on learnability [4, 78]. In this section, we first consider the Rademacher complexity of the family $\mathcal{F}_Q$ of functions on the domain Ω,

$$\mathrm{Rad}_S(\mathcal{F}_Q) = \mathbb{E}\Big(\sup_{f\in\mathcal{F}_Q} \frac{1}{S}\sum_{i=1}^{S} \xi_i f(\mathbf{x}_i)\Big), \tag{5.1}$$

where $\mathbf{x}_i, 1 \le i \le S$, are samples in the domain Ω, $\xi_i \in \{-1, 1\}, 1 \le i \le S$, are i.i.d. Rademacher random variables with $P(\xi_i = 1) = P(\xi_i = -1) = 1/2$, and $\mathcal{F}_Q$, $Q \ge 0$, contains all functions on the domain Ω with their Barron norms bounded by Q, see (4.4). In Section 5.1, we derive an upper bound for the Rademacher complexity $\mathrm{Rad}_S(\mathcal{F}_Q)$ that scales with the inverse square root of the sample size S and the square root of the logarithm of the graph order N; see Theorem 5.1 and [2, 17] for similar results in the standard neural network setting.

Let μ be a probability measure on Ω, and $\mathbf{x}_i \in \Omega, 1 \le i \le S$, have entries being i.i.d. random variables with probability measure μ, and define

$$\Phi(\mathbf{X}) = \sup_{f\in\mathcal{F}_Q} \Big| \int_\Omega f(\mathbf{x}) d\mu(\mathbf{x}) - \frac{1}{S}\sum_{i=1}^{S} f(\mathbf{x}_i)\Big|, \tag{5.2}$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_S)$ and $Q \ge 0$. A natural question concerning the graph Barron space is whether signals lying in a bounded subset of this space can be efficiently learned from data. In Section 5.2, we address this question by considering the error measure in (5.2) for signals in the ball $\mathcal{F}_Q$ under the noiseless random sampling scenario. In particular, based on the estimate of the Rademacher complexity $\mathrm{Rad}_S(\mathcal{F}_Q)$ in Theorem 5.1, we establish an upper bound for the error measure $\Phi(X)$ in Theorem 5.4 with high probability. This error bound scales linearly with the bound Q and inversely with the square root of the sample size S, with a constant that depends on the square root of the logarithm of the underlying graph size N and the mismatch probability δ. From Theorem 5.4, we may conclude that functions in the graph Barron space can be **efficiently** learned from their noiseless sampling data with high probability.

5.1 Radmacher Complexity

In the following theorem, we provide an upper bound for the Rademacher complexity $\mathrm{Rad}_S(\mathcal{F}_Q)$ defined in (5.1).

Theorem 5.1 *Let $\mathcal{B}$ be the Barron space in (3.7) with the norm $\|\cdot\|$ in (2.17) replaced by the standard ℓ^∞-norm $\|\cdot\|_\infty$, and the convolution norm $\|\cdot\|_{\mathrm{co}}$ satisfying the additional assumption (2.21). For any $\mathbf{x}_i \in \Omega, 1 \le i \le S$, define the Rademacher complexity $\mathrm{Rad}_S(\mathcal{F}_Q)$ of the family $\mathcal{F}_Q$, $Q \ge 0$, as in (5.1). Then*

$$\mathrm{Rad}_S(\mathcal{F}_Q) \le 2\Big(D_0 D_2\sqrt{2\ln(2N)} + \sqrt{2\ln 2}\Big) Q S^{-1/2}, \tag{5.3}$$

where D_0 and D_2 are the constants in (2.14) and (2.21) to measure the uniform bunded of the domain Ω and Lipschitz bound of the graph convolution operation, respectively.

We follow the argument used in [17, Theorem 3] and for the completeness of this paper, we include a sketch of the proof. To prove Theorem 5.1. we recall the contraction lemma and Massart lemma, where $\xi_i, 1 \le i \le S$, are i.i.d. Rademacher random variables [57].

Lemma 5.2 *Let $\mathcal{K}$ be a family of functions on Ω and $\mathbf{x}_i \in \Omega, 1 \le i \le S$. Then*

$$\mathbb{E} \sup_{g\in\mathcal{K}} \sum_{i=1}^{S} \xi_i \sigma\big(g(\mathbf{x}_i)\big) \le \mathbb{E} \sup_{g\in\mathcal{K}} \sum_{i=1}^{S} \xi_i g(\mathbf{x}_i). \tag{5.4}$$

Lemma 5.3 *Let $\mathcal{T} \subset \mathbb{R}^S$ be a finite set with its cardinality denoted by $\#\mathcal{T}$. Then*

$$\mathbb{E}\Big(\max_{\mathbf{t}\in\mathcal{T}} \sum_{i=1}^{S} \xi_i t_i \Big) \le \sqrt{2 \ln \#\mathcal{T}} \max_{\mathbf{t}\in\mathcal{T}} \|\mathbf{t}\|_2,$$

where $\mathbf{t} = [t_1, \ldots, t_S]^T \in \mathcal{T}$.

Now we are ready to prove Theorem 5.1.

Proof of Theorem 5.1 First we show that

$$\sup_{f\in\mathcal{F}_Q} \sum_{i=1}^{S} \xi_i f(\mathbf{x}_i) \le Q \sup_{(\mathbf{b},\mathbf{c})\in\mathbb{T}} \Big\| \sum_{i=1}^{S} \xi_i \sigma(\mathbf{b} * \mathbf{x}_i + \mathbf{c}) \Big\|_\infty \tag{5.5}$$

hold for all $\xi_i \in \{-1, 1\}$ and $\mathbf{x}_i \in \Omega, 1 \le i \le S$.

By Lemma 3.3, there exists a probability measure $\hat{\rho}$ on $\mathbb{S} \times \mathbb{T}$ for any $f \in \mathcal{B}$ such that

$$f(\mathbf{x}) = \|f\|_{\mathcal{B}} \int_{\mathbb{S}\times\mathbb{T}} \mathbf{a}^T \sigma(\mathbf{b} * \mathbf{x} + \mathbf{c}) \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}).$$

Then

$$\begin{aligned}
\sum_{i=1}^{S} \epsilon_i f(\mathbf{x}_i) &= \|f\|_{\mathcal{B}} \int_{\mathbb{S}\times\mathbb{T}} \mathbf{a}^T \Big(\sum_{i=1}^{S} \xi_i \sigma(\mathbf{b} * \mathbf{x}_i + \mathbf{c}) \Big) \hat{\rho}(d\mathbf{a}, d\mathbf{b}, d\mathbf{c}) \\
&\le \|f\|_{\mathcal{B}} \sup_{\mathbf{a}\in\mathbb{S},(\mathbf{b},\mathbf{c})\in\mathbb{T}} \mathbf{a}^T \Big(\sum_{i=1}^{S} \xi_i \sigma(\mathbf{b} * \mathbf{x}_i + \mathbf{c}) \Big) \\
&\le \|f\|_{\mathcal{B}} \sup_{(\mathbf{b},\mathbf{c})\in\mathbb{T}} \Big\| \sum_{i=1}^{S} \xi_i \sigma(\mathbf{b} * \mathbf{x}_i + \mathbf{c}) \Big\|_\infty.
\end{aligned}$$

Taking supremum over all $f \in \mathcal{F}_Q$ in the above estimate proves (5.5).

Next we use (5.5) and apply the contraction lemma to show that

$$\mathbb{E}\sup_{(\mathbf{b},\mathbf{c})\in\mathbb{T}}\Big\|\sum_{i=1}^{S}\xi_i\sigma(\mathbf{b}*\mathbf{x}_i+\mathbf{c})\Big\|_\infty\le 2D_2\mathbb{E}\Big\|\sum_{i=1}^{S}\xi_i\mathbf{x}_i\Big\|_\infty+2\mathbb{E}\Big|\sum_{i=1}^{S}\xi_i\Big|. \tag{5.6}$$

For a graph signal $\mathbf{y}\in\mathbb{R}^V$, we denote its value at vertex $v\in V$ by $\mathbf{y}(v)$, $v\in V$. Observe that

$$\begin{aligned}\sup_{(\mathbf{b},\mathbf{c})\in\mathbb{T}}\Big\|\sum_{i=1}^{S}\xi_i\sigma(\mathbf{b}*\mathbf{x}_i+\mathbf{c})\Big\|_\infty&=\sup_{(\mathbf{b},\mathbf{c})\in\mathbb{T}}\max_{1\le n\le N}\max_{\xi\in\{-1,1\}}\Big(\xi\sum_{i=1}^{S}\xi_i\sigma(\mathbf{b}*\mathbf{x}_i+\mathbf{c})\Big)(v)\\&\le\sum_{\xi\in\{-1,1\}}\max_{1\le n\le N}\sup_{(\mathbf{b},\mathbf{c})\in\mathbb{T}}\Big(\xi\sum_{i=1}^{S}\xi_i\sigma(\mathbf{b}*\mathbf{x}_i+\mathbf{c})\Big)(v),\end{aligned}$$

where the inequality holds as

$$\sup_{(\mathbf{b},\mathbf{c})\in\mathbb{T}}\Big(\pm\sum_{i=1}^{S}\xi_i\sigma(\mathbf{b}*\mathbf{x}_i+\mathbf{c})\Big)(v)\ge\sup_{\|\mathbf{c}\|_\infty=1}\pm\sum_{i=1}^{S}\xi_i\sigma(\mathbf{c}(v))\ge 0,\ v\in V.$$

This together with Lemma 5.2 with $\mathcal{K}=\{(\mathbf{b}*\mathbf{x}+\mathbf{c})(v),\ v\in V,(\mathbf{b},\mathbf{c})\in\mathbb{T}\}$ implies that

$$\begin{aligned}\mathbb{E}\sup_{(\mathbf{b},\mathbf{c})\in\mathbb{T}}\Big\|\sum_{i=1}^{S}\xi_i\sigma(\mathbf{b}*\mathbf{x}_i+\mathbf{c})\Big\|_\infty&\le 2\mathbb{E}\max_{v\in V}\sup_{(\mathbf{b},\mathbf{c})\in\mathbb{T}}\sum_{i=1}^{S}\xi_i\sigma\big((\mathbf{b}*\mathbf{x}_i+\mathbf{c})(v)\big)\\&\le 2\mathbb{E}\max_{v\in V}\sup_{(\mathbf{b},\mathbf{c})\in\mathbb{T}}\sum_{i=1}^{S}\xi_i(\mathbf{b}*\mathbf{x}_i+\mathbf{c})(v)\\&\le 2\mathbb{E}\sup_{(\mathbf{b},\mathbf{c})\in\mathbb{T}}\Big\|\mathbf{b}*\Big(\sum_{i=1}^{S}\xi_i\mathbf{x}_i\Big)+\Big(\sum_{i=1}^{S}\xi_i\Big)\mathbf{c}\Big\|_\infty.\end{aligned}$$

Combining the above estimate with (2.21) and the definition of the set $\mathbb{T}$, we complete the proof of (5.6).

Finally we apply (5.6) and use the Massart Lemma to prove (5.3).

Observe that

$$\Big\|\sum_{i=1}^{S}\xi_i\mathbf{x}_i\Big\|_\infty=\max_{v\in V}\max\Big\{\sum_{i=1}^{S}\xi_i\mathbf{x}_i(v),\sum_{i=1}^{S}\xi_i(-\mathbf{x}_i)(v)\Big\},$$

where $\mathbf{x}_i(v)$, $v\in V$, are the value at vertex v of the graph signal $\mathbf{x}_i\in\mathbb{R}^V$. Applying Lemma 5.3 with $\mathcal{T}=\{[\mathbf{x}_1(v),\dots,\mathbf{x}_S(v)]^T,v\in V\}\cup\{-[\mathbf{x}_1(v),\dots,\mathbf{x}_S(v)]^T,v\in$

$V\}$, we conclude that

$$\mathbb{E}\Big\|\sum_{i=1}^{S}\xi_i\mathbf{x}_i\Big\|_\infty \le \sqrt{2\ln(2N)}\sup_{v\in V}\Big(\sum_{i=1}^{S}(\mathbf{x}_i(v))^2\Big)^{1/2}.$$

This together with (2.14) implies that

$$\mathbb{E}\Big\|\sum_{i=1}^{S}\xi_i\mathbf{x}_i\Big\|_\infty \le D_0\sqrt{2\ln(2N)}S^{1/2}. \tag{5.7}$$

Let $\mathbf{1}_S$ be the S-dimensional vector with all components taking value one. Applying Lemma 5.3 with $\mathcal{T} = \{-\mathbf{1}_S, \mathbf{1}_S\}$ gives

$$\mathbb{E}\Big|\sum_{i=1}^{S}\xi_i\Big| \le \sqrt{2\ln 2}S^{1/2}. \tag{5.8}$$

Combining (5.6), (5.7) and (5.8), we complete the proof of the desired estimate (5.3) on the Rademacher complexity. □

5.2 Learnability by GCNNs

In the following theorem, we investigate the learnability of signals in the graph Barron space $\mathcal{B}$ by shallow GCNNs and establish an upper bound for the error measure $\Phi(\mathbf{X})$ defined in (5.2).

Theorem 5.4 *Let $Q \ge 0$, the probability measure μ on Ω, i.i.d. random variables $\mathbf{x}_i \in \Omega, 1 \le i \le S$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_S)$ and let $\Phi(\mathbf{X})$ be as in (5.9). Then for any $\delta \in (0, 1/2)$, the error estimate*

$$\Phi(\mathbf{X}) \le \Big(4D_0D_2\sqrt{2\ln(2N)} + 4\sqrt{2\ln 2} + \sqrt{2\ln(1/\delta)}\Big)QS^{-1/2} \tag{5.9}$$

hold with probability at least $1-\delta$, where D_0 and D_2 are the constants in (2.14) and (2.21) respectively.

Proof We follow the procedure outlined in [4, Theorem 8], incorporating the new estimate of the Rademacher complexity $\mathrm{Rad}_S(\mathcal{F}_Q)$ from Theorem 5.1. For the sake of completeness, we include a brief proof here.

Set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_S)$ and let $\Phi(\mathbf{X})$ be as in (5.2). By the symmetry of the set $\mathcal{F}_Q$, we have

$$\Phi(\mathbf{X}) = \sup_{f\in\mathcal{F}_Q}\int_\Omega f(\mathbf{x})d\mu(\mathbf{x}) - \frac{1}{S}\sum_{i=1}^{S}f(\mathbf{x}_i).$$

By the reproducing kernel property (3.12) for the Barron space, we have

$$|\Phi(\mathbf{X})) - \Phi(\mathbf{X}_i')| \le S^{-1}|f(\mathbf{x}_i) - f(\mathbf{x}_i')| \le 2QS^{-1},\ 1 \le i \le S,$$

where for $1 \le i \le S$, $\mathbf{X}$ and $\mathbf{X}_i'$ share the same components except that their i-th components are $\mathbf{x}_i$ and $\mathbf{x}_i'$ respectively. Then applying McDiarmid's inequality, we obtain

$$\Phi(\mathbf{X}) \le \mathbb{E}_{\mathbf{X}}\Phi(\mathbf{X}) + \sqrt{2\ln(1/\delta)}QS^{-1/2}$$

with probability at least $1-\delta$. Then by Theorem 5.1, it suffices to prove

$$\mathbb{E}_{\mathbf{X}}\Phi(\mathbf{X}) \le 2\mathrm{Rad}_S(\mathcal{F}_Q). \tag{5.10}$$

Let us draw a second sample $\mathbf{x}_1', \ldots, \mathbf{x}_S'$ according to probability measure μ, $\xi_i \in \{-1, 1\}$, $1 \le i \le S$ be i.i.d. Rademacher random variables with $P(\xi_i = 1) = P(\xi_i = -1) = 1/2$, and set $\Xi = (\xi_1, \ldots, \xi_S)$. Then

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}}\Phi(\mathbf{X}) &= \mathbb{E}_{\mathbf{X}} \sup_{f\in\mathcal{F}_Q} \mathbb{E}_{\mathbf{X}'} \frac{1}{S}\sum_{i=1}^{S}\big(f(\mathbf{x}_i') - f(\mathbf{x}_i)\big) \\
&\le \mathbb{E}_{\mathbf{X}}\mathbb{E}_{\mathbf{X}'} \sup_{f\in\mathcal{F}_Q} \frac{1}{S}\sum_{i=1}^{S}\big(f(\mathbf{x}_i') - f(\mathbf{x}_i)\big) \\
&= \mathbb{E}_{\Xi}\mathbb{E}_{\mathbf{X}}\mathbb{E}_{\mathbf{X}'} \sup_{f\in\mathcal{F}_Q} \frac{1}{S}\sum_{i=1}^{S}\xi_i\big(f(\mathbf{x}_i') - f(\mathbf{x}_i)\big) \\
&\le \mathbb{E}_{\Xi}\mathbb{E}_{\mathbf{X}} \sup_{f\in\mathcal{F}_Q} \frac{1}{S}\sum_{i=1}^{S}\xi_i f(\mathbf{x}_i) + \mathbb{E}_{\Xi}\mathbb{E}_{\mathbf{X}'} \sup_{f\in\mathcal{F}_Q} \frac{1}{S}\sum_{i=1}^{S}(-\xi_i) f(\mathbf{x}_i') = 2\mathrm{Rad}_S(\mathcal{F}_Q).
\end{aligned}$$

This proves (5.10) and then completes the proof of Theorem 5.4. □

6 Numerical Simulations

In this section, we consider both synthetic and real data on the underlying undirected graph $\mathcal{G} = (V, E)$ of the data set of hourly temperature collected at 32 weather stations in the region of Brest (France) [48, 56, 82]. The temperature data set is of size $32 \times 24 \times 31$, and the weather station graph $\mathcal{G}$ is constructed by the 5 nearest neighboring stations in physical distances. In this section, we use stochastic gradient descent with Nesterov momentum to train shallow GCNNs on the weather station graph $\mathcal{G}$ and demonstrate the approximation performance of shallow GCNNs presented in Theorems 4.1 and 4.3. All experiments and computations are performed using the PyTorch deep learning framework.

Denote the symmetric normalized Laplacian on the weather station graph $\mathcal{G}$ by $\mathbf{L}^{\mathrm{sym}} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$, where $\mathbf{A}$ and $\mathbf{D}$ are the adjacency and degree matrices of the graph $\mathcal{G}$, respectively. In our simulations, we set $N = 32$, let the fundamental domain Ω of the GCNN contain all graph signals $\mathbf{x} = (x(i))_{i\in V}$ with entries contained in $[-1, 1]$, i.e., $-1 \le x(i) \le 1$ for all $i \in V$, and we use

$$\mathcal{W}_L = \Big\{ \sum_{l=0}^{L} b(l)(\mathbf{L}^{\mathrm{sym}})^l,\ b(0), \ldots, b(L) \in \mathbb{R} \Big\}$$

in (2.13) as the convolution space.

Given input graph signals $\mathbf{x}_i \in \mathbb{R}^N$, $1 \le i \le S$, and output values $y_i = f(\mathbf{x}_i)$, $1 \le i \le S$ of a function f on the domain Ω, we use stochastic gradient descent with Nesterov momentum $\mu = 0.9$, abbreviated to SGDM, as the optimization strategy to learn the parameters $\mathbf{\Theta}$ of the desired GCNN to approximate the function f, see Algorithm 1. The loss function in the SGDM is the conventional relative mean square error (RMSE),

$$F(\mathbf{\Theta}) = \frac{\sum_{i=1}^S (y_i - f_M(\mathbf{x}_i, \mathbf{\Theta}))^2}{\|\mathbf{y}\|_2^2}, \tag{6.1}$$

where $\mathbf{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$ with $\boldsymbol{\theta}_m = (\mathbf{a}_m, \mathbf{b}_m, \mathbf{c}_m) \in \mathbb{R}^N \times \mathbb{R}^{L+1} \times \mathbb{R}^N$ and $\mathbf{b}_m = [b_m(0), \ldots, b_m(L)]^T \in \mathbb{R}^{L+1}$, $1 \le m \le M$, $\|\mathbf{y}\|_2 = (\sum_{i=1}^S y_i^2)^{1/2}$, and

$$f_M(\mathbf{x}_i, \mathbf{\Theta}) = \frac{1}{M} \sum_{m=1}^M \mathbf{a}_m^T \sigma\Big(\sum_{l=0}^L b_m(l) (\mathbf{L}^{\rm sym})^l \mathbf{x}_i + \mathbf{c}_m \Big), \ 1 \le i \le S. \tag{6.2}$$

For the case that $\mathbf{x}_i$, $1 \le i \le S$, are randomly and independently selected with uniform distribution on $\Omega := [-1, 1]^N$, for large sampling size S we may show that the RMSE $F(\mathbf{\Theta})$ is about the relative approximation error of the GCNN in square norm,

$$F(\mathbf{\Theta}) \approx \frac{\int_\Omega |f(\mathbf{x}) - f_M(\mathbf{x})|^2 d\mathbf{x}}{\int_\Omega |f(\mathbf{x})|^2 d\mathbf{x}}.$$

As the ReLU function σ is not differentiable, we define its approximate derivative $\sigma'_{\rm app}$ by

$$\sigma'_{\rm app}(t) = \frac{\sigma(t+\epsilon) - \sigma(t-\epsilon)}{2\epsilon} = \begin{cases} 1 & \text{if } t \in [\epsilon, +\infty) \\ \frac{t+\epsilon}{2\epsilon} & \text{if } t \in (-\epsilon, \epsilon) \\ 0 & \text{if } t \in (-\infty, -\epsilon], \end{cases}$$

which is also the derivative of the regularization

$$\sigma_\epsilon(t) = \begin{cases} t & \text{if } t \in [\epsilon, +\infty) \\ (t+\epsilon)^2/(4\epsilon) & \text{if } t \in (-\epsilon, \epsilon) \\ 0 & \text{if } t \in (-\infty, -\epsilon] \end{cases}$$

of the ReLU function σ, where $\epsilon = 10^{-5}$ in our simulations. Set

$$\mathbf{b}_m(\mathbf{L}^{\rm sym}) = \sum_{l=0}^L b_m(l) (\mathbf{L}^{\rm sym})^l, \ 1 \le m \le M$$

and

$$z_i(\mathbf{\Theta}) = y_i - f_M(\mathbf{x}_i, \mathbf{\Theta}), \ 1 \le i \le S.$$

Algorithm 1 Stochastic Gradient Descent Algorithm with Nesterov Momentum to learn GCNNs

Inputs: Order L of polynomial filter, number M of neurons, order N of the underlying graph $\mathcal{G}$, number S of samples, Nesterov momentum $\mu = 0.9$, learning rate $\gamma = 0.003$, number of iteration Iter, the symmetrically normalized Laplacian $\mathbf{L}^{\rm sym}$, input signals $\mathbf{x}_i \in \mathbb{R}^N, 1 \le i \le S$, and outputs $y_i = f(\mathbf{x}_i), 1 \le i \le S$ of the function f to be learnt.

Initial: $\mathbf{\Theta} \leftarrow \mathbf{0}$ and $\mathbf{Temp} \leftarrow \nabla F_{\rm app}(\mathbf{\Theta})$.

Iteration:

for $n = 1$

$\mathbf{Grad} \leftarrow \nabla F_{\rm app}(\mathbf{\Theta})$

$\mathbf{Temp} \leftarrow \mu * \mathbf{Temp} + \mathbf{Grad}$

$\mathbf{\Theta} \leftarrow \mathbf{\Theta} - \gamma * \mathbf{Temp}$

$n \leftarrow n + 1$

stop if $n >$ Iter

Output: $\mathbf{\Theta}_{\rm Iter}$.

In the SGDM, we use the approximate gradient $\nabla F_{\rm app}$ of the loss function F:

$$\frac{\partial F_{\rm app}(\mathbf{\Theta})}{\partial \mathbf{a}_m} = -\frac{2}{M\|\mathbf{y}\|_2^2}\sum_{i=1}^{S} z_i(\mathbf{\Theta})\sigma\big(\mathbf{b}_m(\mathbf{L}^{\rm sym})\mathbf{x}_i + \mathbf{c}_m\big), \tag{6.3}$$

$$\frac{\partial F_{\rm app}(\mathbf{\Theta})}{\partial \mathbf{c}_m} = -\frac{2}{M\|\mathbf{y}\|_2^2}\sum_{i=1}^{S} z_i(\mathbf{\Theta})\mathrm{diag}\big(\sigma'_{\rm app}\big(\mathbf{b}_m(\mathbf{L}^{\rm sym})\mathbf{x}_i + \mathbf{c}_m\big)\big)\mathbf{a}_m, \tag{6.4}$$

and

$$\frac{\partial F_{\rm app}(\mathbf{\Theta})}{\partial \mathbf{b}_m(l)} = -\frac{2}{M\|\mathbf{y}\|_2^2}\sum_{i=1}^{S} z_i(\mathbf{\Theta})\mathbf{a}_m^T\mathrm{diag}\big(\sigma'_{\rm app}\big(\mathbf{b}_m(\mathbf{L}^{\rm sym})\mathbf{x}_i + \mathbf{c}_m\big)\big)(\mathbf{L}^{\rm sym})^l\mathbf{x}_i, \tag{6.5}$$

where $1 \le m \le M$ and $0 \le l \le L$.

In our first simulation, we consider the quadratic function

$$f(\mathbf{x}) = \|\mathbf{B}\mathbf{x}\|_2^2, \ \mathbf{x} \in \Omega, \tag{6.6}$$

where $\mathbf{B} = (b(i,j))_{i,j\in V}$ has zero entries except for $b(i,i), i \in V$ and $b(i,j), (i,j) \in E$ being drawn randomly and independently from the uniform distribution on $[-1, 1]$. To learn the above function f from the SGDM, we assume that the given input graph signals $\mathbf{x}_i, 1 \le i \le S$, are randomly and independently selected from the uniform distribution on $[-1, 1]^N$, and the output values $y_i = f(x_i) = \|\mathbf{B}\mathbf{x}_i\|_2^2, 1 \le i \le S$. Shown in Figure 1 is the performance of the SGDM to learn the function f from its sampling data $(\mathbf{x}_i, f(\mathbf{x}_i)), 1 \le i \le S$. We observe from Figure 1 that the SGDM converges and has better performance when the sampling size S increases. This demonstrates the theoretical result in Theorem 5.4 on higher learnability of functions from their random samples of larger size.

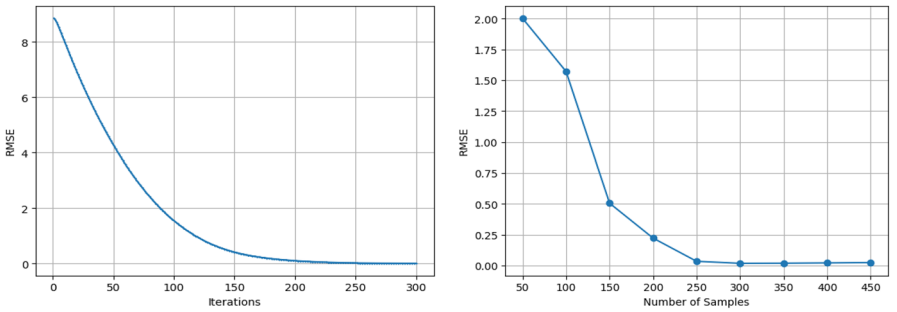

Fig. 1 Plotted on the left is the average RMSE $F(\mathbf{\Theta}_n)$, $1 \le n \le 300$, over 100 trials, where $M = 4$, $L = 5$ and $S = 100$ and $\mathbf{\Theta}_n$, $1 \le n \le 300$, are the parameters in the n-th iteration of the SGDM. The average energy $S^{-1}\|\mathbf{y}\|_2^2$ of the output data over 100 trials is 21.8736, while the uniform bound $\|\mathbf{y}\|_\infty = \sup_{1\le i\le S}|y_i|$ over 100 trials is 6.4467. Presented on the right is the average RMSE $F(\mathbf{\Theta}_{\rm Iter})$ over 100 trials with respect to different sampling size S, where $M = 4$, $L = 5$ and Iter $= 100$

Define the relative uniform approximation error (RUAE) of the GCNN with parameter $\mathbf{\Theta}$ by

$$U(\mathbf{\Theta}) := \frac{\sup_{1\le i\le S}|f(\mathbf{x}_i) - f_M(\mathbf{x}_i, \mathbf{\Theta})|}{\sup_{1\le i\le S}|f(\mathbf{x}_i)|},$$

where f is the original function and $f_M(\mathbf{x}, \mathbf{\Theta})$ is the output of the GCNN given in (6.2). Presented in Figure 2 is how the RMSE and RUAE vary with the number of neurons per vertex. This demonstrates the theoretical result in Theorems 4.1 and 4.3 on the approximation property of GCNNs.

We observe from Figure 2 that increasing the number of neurons at each vertex generally improves the accuracy of the GCNN, as measured by both RMSE and RUAE, as long as the number of iterations in SGDM is not too high. However, when the number of iterations is high (then RMSE and RUAE are low), adding more neurons does not help and may even hurt the performance of the GCNN.

In the second simulation, we consider the real data set of hourly temperature measured in Celsius collected at 32 weather stations in the region of Brest (France) in January 2014. Denote the regional temperature at t_i-th hour of d-th day by $\mathbf{x}_d^{\rm org}(t_i)$, $0 \le i \le 23$, $1 \le d \le 31$. Before we apply GCNNs to learn functions, we pre-process the temperature data set by eliminating the average temperature and rescaling the range to $[-1, 1]$,

$$\mathbf{x}_d(t_i) = B^{-1}\big(\mathbf{x}_d^{\rm org}(t_i) - \mathbf{x}_{\rm ave}^{\rm org}\big),$$

where $\mathbf{x}_{\rm ave}^{\rm org} = (24\times 31)^{-1}\sum_{i=0}^{23}\sum_{d=1}^{31}\mathbf{x}_d^{\rm org}(t_i)$ the average temperature in the region of Brest (France) for January 2014, and B is chosen so that $\mathbf{x}_d(t_i) \in [-1, 1]^{32}$ for all $1 \le d \le 31$ and $0 \le i \le 23$. In particular, we take $B = 10.35$ in our simulation. In the second simulation, we want to learn GCNNs to approximate the squared variance function $f_{\rm sv}$ of next day,

$$f_{\rm sv}(\mathbf{x}_d(t_i)) = \|\mathbf{x}_{d+1}(t_i)\|_2^2 - (\bar{\mathbf{x}}_{d+1}(t_i))^2,\ 1 \le d \le 30, 0 \le i \le 23, \tag{6.7}$$

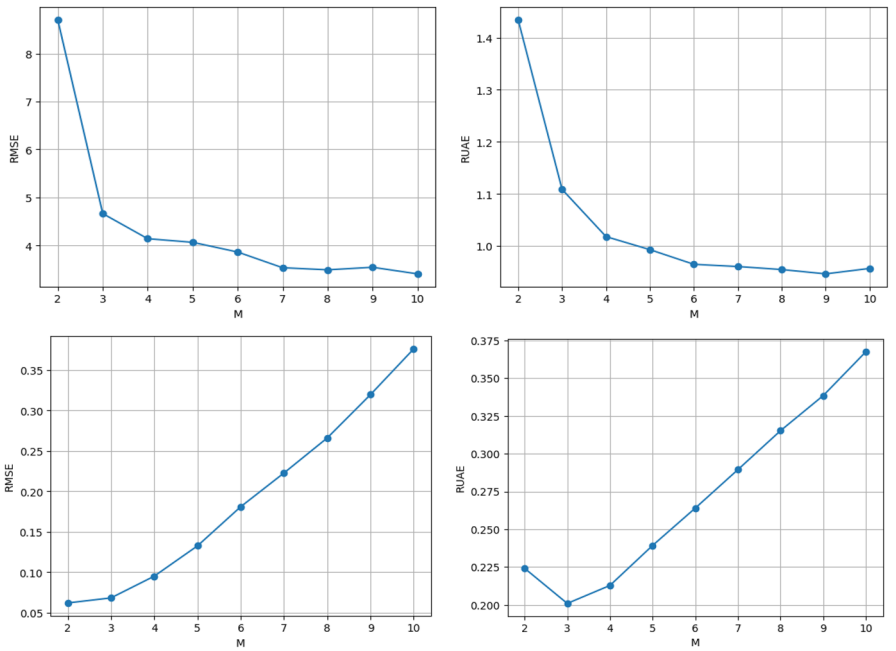

Fig. 2 Plotted are the average RMSE $F(\boldsymbol{\Theta}_{\text{Iter}})$ (left) and RUAE $U(\boldsymbol{\Theta}_{\text{Iter}})$ (right) over 100 trials with respect to different number M of neurons per vertex, where $S = 100$, $L = 5$, and Iter $= 50$ (top left and right) and 200 (bottom left and right) respectively

where $\bar{\mathbf{x}}_{d+1}(t_i)$ is the average pre-processed temperature data of the whole Brest region at t_i-th hour of $(d+1)$-th day.

Learning GCNNs from real-world data is a challenging task. In the second simulation, we try to learn GCNN from about 20% of the weather data set, particularly, $\mathbf{x}_d(t_i)$ and f_{sv}, $0 \le i \le 23$, $d \in \{1, 6, 11, 16, 21, 26\}$, to learn the squared variance function f_{sv} of next day. Shown in Figure 3 is the approximation property of the output of the GCNN obtained from the SGDM, where the relative mean square error (RMSE) and relative uniform approximation error (RUAE) on the **whole** weather data set are defined by

$$\text{WMSE} = \frac{\sum_{i=0}^{23}\sum_{d=1}^{30}|f_{\text{sv}}(\mathbf{x}_d(t_i)) - f_M(\mathbf{x}_d(t_i), \boldsymbol{\Theta}_{\text{Iter}})|^2}{\sum_{i=0}^{23}\sum_{d=1}^{30}|f_{\text{sv}}(\mathbf{x}_d(t_i))|^2}$$

and on the right is the uniform error

$$\text{WUAE} = \frac{\sup_{1\le d\le 30, 0\le i\le 23}|f_{\text{sv}}(\mathbf{x}_d(t_i)) - f_M(\mathbf{x}_d(t_i), \boldsymbol{\Theta}_{\text{Iter}})|}{\sup_{1\le d\le 30, 0\le i\le 23}|f_{\text{sv}}(\mathbf{x}_d(t_i)|},$$

where $\boldsymbol{\Theta}_{\text{Iter}}$ is the output parameter of the SGDM. Comparing with the approximation of the quadratic function f in (6.6) with the squared variance function f_{sv} in (6.7) by GCNNs, the number of neurons at each vertex has a positive impact on the accuracy of

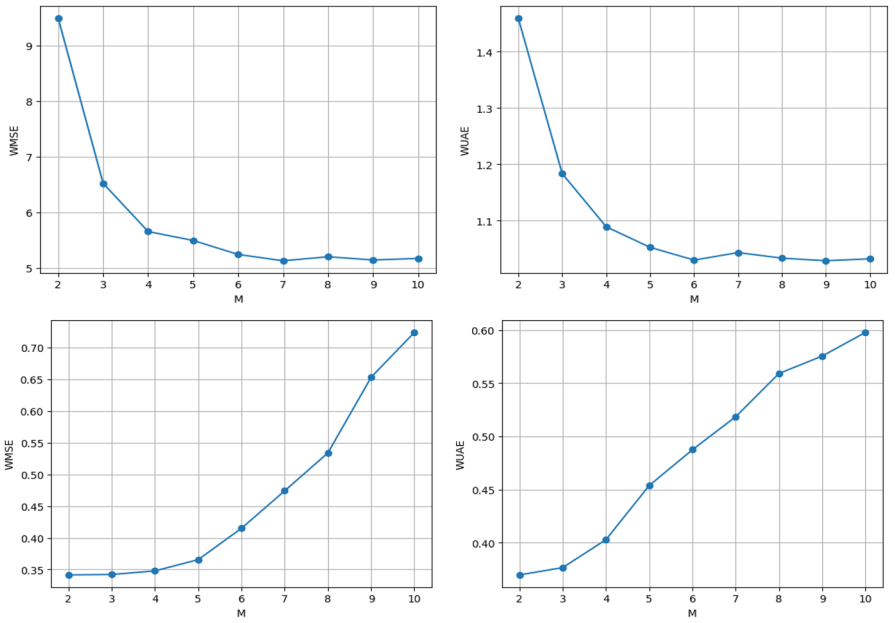

Fig. 3 Plotted are the average WMSE (left) and WUAE (right) over 100 trials with respect to different number M of neurons per vertex, where $S = 218 = 24 \times 12$, $L = 5$, and Iter $= 30$ (top left and right) and 100 (bottom left and right) respectively

the GCNN, when the number of iterations in SGDM is not too high. However, when the number of iterations is high, adding more neurons does not improve and may even degrade the performance of the GCNN. We hypothesize that this is because the GCNN becomes overfitted to the training data and loses its ability to generalize to new data.

From the approximation property presented by Figures 2 and 3, we observe that there is a trade-off between the number M of neurons and the number Iter of iterations in the SGDM that needs to be carefully balanced to achieve the optimal performance of the GCNNs.

7 Conclusion and Discussions

In this paper, we introduce a Barron space $\mathcal{B}$ associated with two-layer GCNNs in the spectral convolution setting and show that functions in the Barron space can be well approximated by outputs of two-layer GCNNs without suffering from the curse of dimensionality.

For a graph filter $\mathbf{G} = (G(i, j))_{i,j\in V}$, define its geodesic-width $\omega(\mathbf{G})$ by the smallest nonnegative integer such that $G(i, j) = 0$ for all $i, j \in V$ with the geodesic distance $\rho(i, j)$ between vertices i and j is strictly larger than $\omega(\mathbf{G})$. Denote the set of all matrices with their geodesic-width no larger than L by $\mathcal{M}_L$. In the spatial approach to define graph convolution, a localized matrix operation $\mathbf{x} \longmapsto \mathbf{Hx}$ associated with some matrix $\mathbf{H} \in \mathcal{M}_L$ is applied instead of the spectral convolution $\mathbf{x} \longmapsto \mathbf{b} * \mathbf{x}$

associated with a graph signal $\mathbf{b}$. In particular, given a graph signal $\mathbf{b}$ in the convolution space $\mathcal{W}$, we can find a polynomial filter $\mathbf{H} = h(\mathbf{S}_1, \ldots, \mathbf{S}_K)$ in some $\mathcal{M}_L$ such that the corresponding spectral convolution can be implemented by polynomial filtering procedure, i.e., $\mathbf{b} * \mathbf{x} = \mathbf{H}\mathbf{x}$ holds for any graph signal $\mathbf{x}$, where L is the degree of the multivariate polynomial h. Comparing with the spectral convolution setting, the convolution in the spatial setting has much more parameters to learn, as the convolution space $\mathcal{W}$ has $\dim \mathcal{W} \le N$, while the convolution space $\mathcal{M}_L$ in the spatial setting has dimension bounded below by N and above by N^2, i.e., $N \le \dim \mathcal{M}_L \le \dim \mathcal{M}_D = N^2$, where $D = \max_{i,j\in V} \rho(i, j)$ is the diameter of the graph $\mathcal{G}$.

In the spatial convolution setting, the output of a two-layer GCNN is given by

$$f_M(x, \mathbf{\Theta}) = \frac{1}{M} \sum_{m=1}^{M} \mathbf{a}_m^T \sigma(\mathbf{H}_m \mathbf{x} + \mathbf{c}_m)$$

where $\mathbf{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$ and $\boldsymbol{\theta}_m = (\mathbf{a}_m, \mathbf{H}_m, \mathbf{c}_m) \in \mathbb{R}^N \times \mathcal{M}_L \times \mathbb{R}^N$, $1 \le m \le M$, see [8, 33] for $L = 1$. With appropriate convolution norm for matrices in $\mathcal{M}_L$, we can define a Barron space associated with two-layer GCNNs in the spatial convolution setting for functions with the following representation

$$f(\mathbf{x}) = \int_{\mathbb{R}^N \times \mathcal{M}_L \times \mathbb{R}^N} \mathbf{a}^T \sigma(\mathbf{H}\mathbf{x} + \mathbf{c}) \rho(d\mathbf{a}, d\mathbf{H}, d\mathbf{c}), \ \mathbf{x} \in \Omega,$$

and show that functions in the Barron space can be approximated by two-layer GCNNs, where ρ is a probability measure on $\mathbb{R}^N \times \mathcal{M}_L \times \mathbb{R}^N$, cf. (3.5), (3.6) and (3.7). Following the argument in Remark 4.9, we can show that the Barron space in the spatial convolution setting is not dense in the space of continuous functions on the unit ball, provided that the maximal geodesic-distance L chosen for the convolution space $\mathcal{M}_L$ is strictly less than the diameter of the graph $\mathcal{G}$.

Acknowledgements The authors acknowledge the reviewers for their constructive comments for the improvement of the paper, and also the assistance of Microsoft Copilot and DeepSeek to provide phrasing refinements and typesetting suggestions in the revision.

References

1. Akoglu, L., Tong, H., Koutra, D.: Graph based anomaly detection and description: a survey. Data Min. Knowl. Discov. **29**, 626–688 (2015)

2. Bach, F.: Breaking the curse of dimensionality with convex neural networks. J. Mach. Learn. Res. **18**, 1–53 (2017)
3. Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Trans. Inform. Theory **39**, 930–945 (1993)
4. Bartlett, P.L., Mendelson, S.: Rademacher and Gaussian complexities: risk bounds and structural results. J. Mach. Learn. Res. **3**, 463–482 (2002)
5. Bartolucci, F., De Vito, E., Rosasco, L., Vigogna, S.: Understanding neural networks with reproducing kernel Banach spaces. Appl. Comput. Harmon. Anal. **62**, 194–236 (2023)
6. Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A., Vandergheynst, P.: Geometric deep learning: going beyond Euclidean data. IEEE Signal Process. Mag. **34**, 18–42 (2017)
7. Brüel Gabrielsson, R.: Universal function approximation on graphs. In: Proceeding of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), p. 11 (2020)
8. Bruna, J., Zaremba, W., Szlam, A., LeCun, Y.: Spectral networks and locally connected networks on graphs. In: Proceeding of the International Conference on Learning Representations (ICLR2014), p. 14 (2014)
9. Chen, Y., Cheng, C., Sun, Q.: Graph Fourier transform based on singular value decomposition of directed Laplacian. Sampl. Theory Signal Process. Data Anal. **21**(2), 24 (2023)
10. Cheney, E.W., Light, W.A.: A course in approximation theory graduate studies in mathematics. Amer. Math. Soc., 101 (2000)
11. Cheng, C., Chen, Y., Lee, Y.J., Sun, Q.: SVD-based graph Fourier transforms on directed product graphs. IEEE Trans. Signal Inform. Process. Netw. **9**, 531–541 (2023)
12. Cheung, M., Shi, J., Wright, O., Jiang, L.Y., Liu, X., Moura, J.M.F.: Graph signal processing and deep learning: convolution, pooling, and topology. IEEE Signal Process. Mag. **37**, 139–149 (2020)
13. Chong, C.-Y., Kumar, S.P.: Sensor networks: evolution, opportunities, and challenges. Proc. IEEE **91**, 1247–1256 (2003)
14. Chung, F.R.K.: Spectral Graph Theory. CBMS regional conference series in mathematics, vol. 92. Amer. Math. Soc, Providence, RI (1997)
15. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. In: Proceeding of the 30th Conference on Neural Information Processing Systems (NeuIPS 2016), p. 9 (2016)
16. Dong, X., Thanou, D., Toni, L., Bronstein, M., Frossard, P.: Graph signal processing for machine learning: a review and new perspectives. IEEE Signal Process. Mag. **37**, 117–127 (2020)
17. E, W., Ma, C., Wu, L.: The Barron space and the flow-induced function spaces for neural network models. Constr. Approx. **55**, 369–406 (2022)
18. E, W., Ma, C., Wu, L., Wojtowytsch, S.: Towards a mathematical understanding of neural network-based machine learning: what we know and what we don't. CSIAM Trans. Appl. Math **1**, 561–615 (2020)
19. Weinan, E., Wojtowytsch, S.: Representation formulas and pointwise properties for Barron functions. Calc. Var. Partial. Differ. Equ. **61**(2), 46 (2022)
20. Emirov, N., Cheng, C., Jiang, J., Sun, Q.: Polynomial graph filters of multiple shifts and distributed implementation of inverse filtering. Sampl. Theory Signal Process. Data Anal. **20**(1), 2 (2022)
21. Emirov, N., Song, G., Sun, Q.: A divide-and-conquer algorithm for distributed optimization on networks. Appl. Comput. Harmon. Anal. **70**, 101623 (2024)
22. Fasshauer, G.E., Hickernell, F.J., Ye, Q.: Solving support vector machines in reproducing kernel Banach spaces with positive definite functions. Appl. Comput. Harmon. Anal. **38**, 115–139 (2015)
23. Gavili, A., Zhang, X.-P.: On the shift operator, graph frequency, and optimal filtering in graph signal processing. IEEE Trans. Signal Process. **65**, 6303–6318 (2017)
24. Ghojogh, B., Ghodsi, A., Karray, F., Crowley, M.: Reproducing kernel Hilbert space, Mercer's theorem, eigenfunctions, Nyström Method, and use of kernels in machine learning: tutorial and survey (2021). arXiv:2106.08443
25. Giannakis, G.B., Ling, Q., Mateos, G., Schizas, I.D., Zhu, H.: Decentralized learning for wireless communications and networking. In: Glowinski, R., Osher, S.J., Yin, W. (eds.) Splitting Methods in Communication, Imaging, Science, and Engineering, pp. 461–497. Springer International Publishing, Cham (2016)
26. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., Chen, T.: Recent advances in convolutional neural networks. Pattern Recognit. **77**, 354–377 (2018)
27. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (2012)

28. Isufi, E., Gama, F., Shuman, D.I., Segarra, S.: Graph filters for signal processing and machine learning on graphs. IEEE Trans. Signal Process. **72**, 4745–4781 (2024)
29. Jiang, J., Cheng, C., Sun, Q.: Nonsubsampled graph filter banks: theory and distributed algorithms. IEEE Trans. Signal Process. **67**, 3938–3953 (2019)
30. Kekatos, V., Giannakis, G.B.: Distributed robust power system state estimation. IEEE Trans. Power Syst. **28**, 1617–1626 (2013)
31. Keriven, N., Bietti, A., Vaiter, S.: On the universality of graph neural networks on large random graphs. In: Proceeding of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), p. 12 (2021)
32. Keriven, N., Peyré, G.: Universal invariant and equivariant graph neural networks. In: The Proceeding of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), p. 10 (2019)
33. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional network. In: Proceeding of ICLR 2017 (Poster), p. 14 (2017)
34. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Commun. ACM **60**, 84–90 (2017). (**The original version in NeuIPS 2012**)
35. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Handwritten digit recognition with a back-propagation network. In: Proceeding of the Advances in Neural Information Processing Systems (NeuIPS 1989), pp. 396–404 (1989)
36. Li, R., Wang, S., Zhu, F., Huang, J.: Adaptive graph convolutional neural networks. In: Proceeding of Thirty-Second AAAI Conference on. Artif. Intell. **32**, 3546–3553 (2018)
37. Li, Z., Liu, F., Yang, W., Peng, S., Zhou, J.: A survey of convolutional neural networks: analysis, applications, and prospects. IEEE Trans. Neural Netw. Learn. Syst. **33**, 6999–7019 (2022)
38. Lin, R.R., Zhang, H.Z., Zhang, J.: On reproducing kernel Banach spaces: generic definitions and unified framework of constructions. Acta. Math. Sin.-English Ser **38**, 1459–1483 (2022)
39. Liu, Z., Zhou, J.: Introduction to Graph Neural Networks. Springer, Cham (2020)
40. Mao, T., Shi, Z., Zhou, D.-X.: Approximating functions with multi-features by deep convolutional neural networks. Anal. Appl. (Singap.) **21**, 1–31 (2022)
41. Makhdoumi, A., Ozdaglar, A.: Convergence rate of distributed ADMM over networks. IEEE Trans. Autom. Control **62**, 5082–5095 (2017)
42. Molzahn, D.K., Dörfler, F., Sandberg, H., Low, S.H., Chakrabarti, S., Baldick, R., Lavaei, J.: A survey of distributed optimization and control algorithms for electric power systems. IEEE Trans. Smart Grid **8**, 2941–2962 (2017)
43. Moore, N.S., Cyr, E.C., Ohm, P., Siefert, C.M., Tuminaro, R.S.: Graph neural networks and applied linear algebra. SIAM Rev. **67**, 141–175 (2025)
44. Nashed, M.Z., Sun, Q.: Sampling and reconstruction of signals in a reproducing kernel subspace of $L^p(\mathbb{R}^d)$. J. Funct. Anal. **258**, 2422–2452 (2010)
45. Niepert, M., Ahmed, M., Kutzkov, K.: Learning convolutional neural networks for graphs. In: Proceeding of the 33rd International Conference on Machine Learning (ICML), pp. 2014–2023 (2016)
46. Ortega, A.: Introduction to Graph Signal Processing. Cambridge University Press, Cambridge (2022)
47. Ortega, A., Frossard, P., Kovačević, J., Moura, J.M.F., Vandergheynst, P.: Graph signal processing: overview, challenges, and applications. Proc. IEEE **106**(5), 808–828 (2018)
48. Perraudin, N., Vandergheynst, P.: Stationary signal processing on graphs. IEEE Trans. Signal Process. **65**(13), 3462–3477 (2017)
49. Pinkus, A.: Approximation theory of the MLP model in neural networks. Acta Numer **8**, 143–195 (1999)
50. Prokhorov, Y.V.: Convergence of random processes and limit theorems in probability theory. Theory Probab. Appl. **1**, 157–214 (1956)
51. Rahimi, A., Recht, B.: Random features for large-scale kernel machines. In: Proceeding of Advances in Neural Information Processing Systems 20 (NeuIPS 2007), p. 8 (2007)
52. Ricaud, B., Borgnat, P., Tremblay, N., Gonçalves, P., Vandergheynst, P.: Fourier could be a data scientist: from graph Fourier transform to signal processing on graphs. C R Phys. **20**, 474–488 (2019)
53. Sandryhaila, A., Moura, J.M.F.: Discrete signal processing on graphs. IEEE Trans. Signal Process. **61**, 1644–1656 (2013)
54. Sandryhaila, A., Moura, J.M.F.: Discrete signal processing on graphs: frequency analysis. IEEE Trans. Signal Process. **62**, 3042–3054 (2014)
55. Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, Massachusetts (2002)

56. Segarra, S., Marques, A.G., Leus, G., Ribeiro, A.: Stationary graph processes: parametric power spectral estimation, In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, USA, pp. 4099-4103 (2017)
57. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, Cambridge (2014)
58. Shen, Z., Yang, H., Zhang, S.: Optimal approximation rate of ReLU networks in terms of width and depth. J. Math. Pures Appl. **9**(157), 101–135 (2022)
59. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceeding of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12026–12035 (2019)
60. Shuman, D.I., Narang, S.K., Frossard, P., Ortega, A., Vandergheynst, P.: The emerging field of signal processing on graphs: extending high-dimensional data analysis to networks and other irregular domains. IEEE Signal Process. Mag. **30**, 83–98 (2013)
61. Siegel, J.W., Xu, J.: Characterization of the variation spaces corresponding to shallow neural networks. Constr. Approx. **57**, 1109–1132 (2023)
62. Song, G., Zhang, H., Hickernell, F.J.: Reproducing kernel Banach spaces with the ℓ^1 norm. Appl. Comput. Harmon. Anal. **34**, 96–116 (2013)
63. Spek, L., Heeringa, T.J., Schwenninger, F., Brune, C.: Duality for neural networks through reproducing kernel Banach spaces. Appl. Comput. Harmon. Anal. **78**, 101765 (2025)
64. Stanković, L., Daković, M., Sejdić, E.: Introduction to graph signal processing, In: Vertex-Frequency Analysis of Graph Signals, pp. 3–108. Springer Cham (2019)
65. Steinwart, I., Christmann, A.: Support Vector Machines. Springer-Verlag, New York (2008)
66. Vershynin, R.: On the role of sparsity in compressed sensing and random matrix theory. In: Proceeding of the 3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), pp. 189–192 (2009)
67. Wang, R., Xu, Y.: Representation theorems in banach spaces: minimum norm interpolation, regularized learning and semi-discrete inverse problems. J. Mach. Learn. Res. **22**, 1–65 (2021)
68. Wang, R., Xu, Y., Yan, M.: Hypothesis spaces for deep learning. Neural Netw. **193**, 107995 (2026)
69. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)
70. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Yu, P.S.: A comprehensive survey on graph neural networks. IEEE Trans. Neural Netw. Learn. **32**, 4–24 (2021)
71. Xu, Y.: Sparse machine learning in Banach spaces. Appl. Numer. Math. **187**, 138–157 (2023)
72. Xu, Y., Ye, Q.: Generalized Mercer kernels and reproducing kernel Banach spaces. Mem. Am. Math. Soc **258**, 122 (2019)
73. Yang, W., Zhang, J., Cai, J., Xu, Z.: Shallow graph convolutional network for skeleton-based action recognition. Sensors **21**(2), 452 (2021)
74. Yang, Y., Feng, H., Zhou, D.-X.: On the rates of convergence for learning with convolutional neural networks. SIAM J. Math. Data Sci. **7**, 1755–1772 (2025)
75. Yang, Y., Zhou, D.-X.: Nonparametric regression using over-parameterized shallow ReLU neural networks. J. Mach. Learn. Res. **25**(165), 1–35 (2024)
76. Yang, Y., Zhou, D.-X.: Optimal rates of approximation by shallow ReLU^k neural networks and applications to nonparametric regression. Constr. Approx. **62**, 329–360 (2025)
77. Yick, J., Mukherjee, B., Ghosal, D.: Wireless sensor network survey. Comput. Netw. **52**, 2292–2330 (2008)
78. Yin, D., Kannan, R., Bartlett, P.: Rademacher complexity for adversarially robust generalization. In: Proceeding of the 36th International Conference on Machine Learning (ICML), PMLR, vol. 97, pp. 7085–7094 (2019)
79. Zhang, H., Xu, Y., Zhang, J.: Reproducing kernel Banach spaces for machine learning. J. Mach. Learn. Res. **10**, 2741–2775 (2009)
80. Zhang, H., Zhang, J.: Vector-valued reproducing kernel Banach spaces with applications to multi-task learning. J. Complexity **29**, 195–215 (2013)
81. Zhang, S., Tong, H., Xu, J., Maciejewski, R.: Graph convolutional networks: a comprehensive review. Comput. Soc. Netw. **6**(1), 1–23 (2019)
82. Zheng, C., Cheng, C., Sun, Q.: Wiener filters on graphs and distributed implementation. Digit. Signal Process **162**, 105156 (2025)

83. Zhou, D.-X.: Universality of deep convolutional neural networks. Appl. Comput. Harmon. Anal. **48**, 787–794 (2020)
84. Zhou, D.-X.: Theory of deep convolutional neural networks: downsampling. Neural Netw. **124**, 319–327 (2020)
85. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M.: Graph neural networks: a review of methods and applications. AI Open **1**, 57–81 (2020)
86. Zhuang, C., Ma, Q.: Dual graph convolutional networks for graph-based semi-supervised classification. In: Proceeding of 2018 Web Conference, pp. 499–508 (2018)