Deep Distributed Convolutional Neural Networks: Universality

Ding-Xuan Zhou
Department of Mathematics, City University of Hong Kong
Kowloon, Hong Kong
Email: mazhou@cityu.edu.hk

Abstract

Deep learning based on structured deep neural networks has provided powerful applications in various fields. The structures imposed on the deep neural networks are crucial, which makes deep learning essentially different from classical schemes based on fully connected neural networks. One of the commonly used deep neural network structures is generated by convolutions. The produced deep learning algorithms form the family of deep convolutional neural networks. Despite of their power in some practical domains, little is known about the mathematical foundation of deep convolutional neural networks such as universality of approximation. In this paper we propose a family of new structured deep neural networks, deep distributed convolutional neural networks. We show that these deep neural networks have the same order of computational complexity as the deep convolutional neural networks, and we prove their universality of approximation. Some ideas of our analysis are from ridge approximation, wavelets, and learning theory.

Keywords: deep learning, convolutional neural networks, deep distributed convolutional neural networks, universality, filter mask

Mathematics Subject Classification 2000: 68Q32, 68T05

1 Introduction and Main Result

The classical (shallow) neural networks to approximate functions or process data on $\mathbb{R}^d$ take the form

$$f_N(x) = \sum_{k=1}^{N} c_k \sigma([w]_k \cdot x - b_k). \tag{1.1}$$

Here $x := (x_1, x_2, \ldots, x_d)^T \in \mathbb{R}^d$ is the vector of input variables, $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function, N is the number of neurons, and $\{[w]_k \in \mathbb{R}^d, b_k \in \mathbb{R}, c_k \in \mathbb{R}\}$ are

parameters corresponding to weights, biases (or thresholds), and coefficients, with $[w]_i \cdot x$ being the dot product in $\mathbb{R}^d$. Approximation of functions on subsets of $\mathbb{R}^d$ by the shallow neural networks (1.1) was studied well in a large classical literature in the late 1980s, which is described in the survey [23] and references therein. A particular research problem called **universality of approximation** is to consider when a neural network of the form (1.1) can approximate any continuous function on any compact subset of $\mathbb{R}^d$ to an arbitrary accuracy when N is large enough, see [6, 13, 1, 16, 21] and references therein.

Approximation by the neural networks (1.1) has been extended to a setting with multi-layer neural networks in the 1990s. A multi-layer neural network with J hidden layers of neurons $\{h^{(j)} : \mathbb{R}^d \to \mathbb{R}^{d_j}\}_{j=1}^J$ with widths $\{d_j\}$ is iteratively generated as

$$h^{(j)}(x) = \Big(\sigma([w]_{i,j} \cdot h^{(j-1)}(x) - b_i^{(j)})\Big)_{i=1}^{d_j}, \tag{1.2}$$

where $h^{(0)}(x) = x \in \mathbb{R}^d$, $d_0 = d$ and $\{[w]_{i,j} \in \mathbb{R}^{d_{j-1}}, b_i^{(j)} \in \mathbb{R}\}$ are free parameters. The last hidden layer produces output functions of the form $f_N(x) = c \cdot h^{(J)}(x)$ with $c \in \mathbb{R}^{d_J}$. The multi-layer neural network (1.2) is reduced to (1.1) when $J = 1$, and its universality and approximation properties have also been well studied in the literature [13, 21, 3, 4, 5].

A key point to ensure the universality of the neural networks (1.1) or (1.2) is the complete freedom in taking the weights $\{[w]_k\}$ in (1.1) or $\{[w]_{i,j}\}$ in (1.2), and these neural networks are called **fully connected** because of this feature. From the fully connectedness one can easily calculate the number of free parameters of weights: dN in (1.1) and $\sum_{j=1}^J d_j d_{j-1}$ in (1.2), very large when the dimension d is high, which makes these neural networks hard to implement for some practical applications dealing with big data in huge dimensions.

Great progress on artificial intelligence has been made after deep learning was introduced [14]. A basic idea of deep learning is to reduce the computational complexity of the multi-layer neural networks involving too many free parameters by imposing architecture designs and applying error-correction tuning methods in graphics processing units such as backpropagation and stochastic gradient descent for computing feasible and satisfactory solutions [9, 12, 15]. These special structured multi-layer neural networks, called **deep neural networks**, have led to practical success of the scalable deep learning algorithms. There have been many deep architectures proposed for various domains of applications and developments of the corresponding efficient deep learning systems for modelling deep abstractions from big data. One important deep architecture is **deep convolutional neural networks** (DCNNs) which have provided powerful applications in domains like computer vision. Compared with their success in practical applications, very little is known about the

approximation properties of DCNNs [2, 22, 17] and there is a dramatically increasing need of rigorous mathematical foundations of DCNNs. To our best knowledge, the universality of the DCNN has not been proved or disproved.

In this paper we propose a family of new deep neural networks, **deep distributed convolutional neural networks** (DDCNNs), and prove their universality of approximation. DDCNNs generalize the DCNNs by allowing multiple biases for distributed implementations which is motivated by our recent work on distributed learning algorithms for dealing with big data [18, 11, 10]. They are generated by means of the rectified linear unit (ReLU) activation function σ defined by

$$\sigma(u) = \max\{u, 0\} = \begin{cases} u, & \text{if } u \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

The universality of approximation of a DDCNN means that it can approximate any continuous function f on any compact subset of $\mathbb{R}^d$ to an arbitrary accuracy when its depth J is large enough. The DDCNN has a fixed **filter length** $s \in \mathbb{N}$ and is generated by a sequence $\{w^{(j)}\}_{j\in\mathbb{N}}$ of **filter masks** $w^{(j)} = (w_i^{(j)})_{i\in\mathbb{Z}} : \mathbb{Z} \to \mathbb{R}$ supported on $\{0, \ldots, s\}$ meaning that $w_i^{(j)} \neq 0$ only when $i \in \{0, \ldots, s\}$.

Let us describe DCNNs before defining DDCNNs. When the depth is $J \in \mathbb{N}$, a DCNN may take the form of a J-layer neural network (1.2) with $\{d_j := d + js\}_{j=1}^J$ and a special convolutional structure for the weights $\{[w]_{i,j}\}$ generated by the filter masks as

$$[w]_{i,j} = \left(w_{i-k}^{(j)}\right)_{k=1}^{d_{j-1}} \in \mathbb{R}^{d_{j-1}}, \qquad i = 1, \ldots, d_j,\ j = 1, \ldots, J. \tag{1.3}$$

By the support property of the mask $w^{(j)}$, the above convolutional structure can be viewed more explicitly from a Toeplitz type matrix $T_{d_{j-1}}^{w^{(j)}}$. Here the Toeplitz type matrix T_D^w associated with a filter mask $w : \mathbb{Z} \to \mathbb{R}$ supported on $\{0, \ldots, s\}$ and a column number $D \geq s$ is defined to be a $(D + s) \times D$ matrix $(w_{i-k})_{i=1,\ldots,D+s,k=1,\ldots,D}$ given by

$$T_D^w = \begin{bmatrix} w_0 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ w_1 & w_0 & 0 & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ w_s & w_{s-1} & \cdots & w_0 & 0 & \cdots & 0 \\ 0 & w_s & \cdots & w_1 & w_0 & 0\cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & w_s & \cdots & w_1 & w_0 \\ 0 & \cdots & 0 & 0 & w_s & \cdots & w_1 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & w_s \end{bmatrix} \in \mathbb{R}^{(D+s)\times D}. \tag{1.4}$$

Observe that the number d_j of neurons at level j of the DCNN increases as j becomes larger and is given iteratively by $d_1 = d + s$ and $d_{j+1} = d_j + s$ for $j \in \mathbb{N}$. Because of the convolutional structure in (1.3), one only needs to compute

$$(s+1)J + \sum_{j=1}^{J-1} d_j + 2d_J = J\left(d + 1 + \frac{J+5}{2}s\right) + d \tag{1.5}$$

parameters for the DCNN, a large reduction from

$$\sum_{j=1}^{J} d_j d_{j-1} + \sum_{j=1}^{J-1} d_j + 2d_J = J\left(d^2 + d + Js\left(\frac{Js}{3} + d + \frac{1}{2}\right) + \frac{3s}{2} - \frac{s^2}{3}\right) + d$$

parameters for the fully connected network (1.2). Note that the k-th component $\left([w]_{i,j}\right)_k = w_{i-k}^{(j)}$ of the vector $[w]_{i,j}$ in (1.3) satisfies

$$\text{for } 1 \le i \le d_j, 1 \le k \le d_{j-1},\ w_{i-k}^{(j)} \ne 0 \Longrightarrow \max\{1, i-s\} \le k \le \min\{i, d_{j-1}\}. \tag{1.6}$$

The purpose of this paper is to introduce the DDCNN which has the same order of computational complexity as the DCNN. To express this new deep neural network, we take an approach different from the form (1.2) for the DCNN. We define a sequence of function vectors $\left\{\Phi^{(j)}(x)\right\}_{j=1}^{J}$, called **DDCNN input vectors**, as appearing in (1.2) before the action of the activation function σ. Denote the sequence of filter masks supported on $\{0, \ldots, s\}$ as $\mathbf{w} = \{w^{(j)}\}_{j=1}^{J}$.

Definition 1. *The DDCNN input vectors* $\left\{\Phi^{(j)}(x) = \left(\phi_i^{(j)}(x)\right)_{i=1}^{d_j}\right\}_{j=1}^{J}$ *of a DDCNN of depth J is defined for the first layer $\Phi^{(1)}(x)$ by*

$$\phi_i^{(1)}(x) = \sum_{k=1}^{d} w_{i-k}^{(1)} x_k = \sum_{k=\max\{1,i-s\}}^{\min\{d,i\}} w_{i-k}^{(1)} x_k, \qquad 1 \le i \le d_1 = d + s, \tag{1.7}$$

and for the more layers $\Phi^{(j)}(x)$ with $j = 2, \ldots, J$ and $1 \le i \le d_j = d + js$ by

$$\phi_i^{(j)}(x) = \sum_{k=1}^{d_{j-1}} w_{i-k}^{(j)} \sigma\left(\phi_k^{(j-1)}(x) - b_{k,i}^{(j-1)}\right) = \sum_{k=\max\{1,i-s\}}^{\min\{d_{j-1},i\}} w_{i-k}^{(j)} \sigma\left(\phi_k^{(j-1)}(x) - b_{k,i}^{(j-1)}\right), \tag{1.8}$$

where $\left(b_{k,i}^{(j-1)} \in \mathbb{R} :\ k = 1, \ldots, d_{j-1},\ i = 1, \ldots, d_j\right) =: b^{(j-1)}$ *is a bias matrix. The generated* **DDCNN hypothesis space** *is defined in terms of* $\mathbf{w}$ *and the sequence of bias matrices* $\mathbf{b} = (b^{(j)})_{j=1}^{J}$ *with* $b^{(J)} \in \mathbb{R}^{d_J}$ *to be a space of functions as*

$$\mathcal{H}_J^{\mathbf{w},\mathbf{b}} = \left\{\sum_{k=1}^{d_J} c_k \sigma\left(\phi_k^{(J)}(x) - b_k^{(J)}\right) : c_1, \ldots, c_{d_J} \in \mathbb{R}\right\}. \tag{1.9}$$

To see (1.8) more explicitly, we set $b_{k,i}^{(j-1)} = 0$ for $k \notin [\max\{1, i-s\}, \min\{d_{j-1}, i\}]$, and express $b^{(j-1)} \in \mathbb{R}^{d_{j-1} \times d_j}$ to be a band matrix with band width s given by

$$b^{(j-1)} = \begin{bmatrix} b_{1,1}^{(j-1)} & b_{1,2}^{(j-1)} & \cdots b_{1,s+1}^{(j-1)} & 0 & \cdots & \cdots & 0 \\ 0 & b_{2,2}^{(j-1)} & \cdots & b_{2,s+2}^{(j-1)} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & & \ddots & \ddots & \vdots \\ & & & & & & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & b_{d_{j-1},d_{j-1}}^{(j-1)} & \cdots & b_{d_{j-1},d_j}^{(j-1)} \end{bmatrix}. \tag{1.10}$$

Then the iterative relationship (1.8) can be written as

$$\Phi^{(j)}(x) = \left(\left(T_{d_{j-1}}^{w^{(j)}} \right)_{i,\cdot} \sigma \left(\Phi^{(j-1)}(x) - \left(b^{(j-1)} \right)_{\cdot,i} \right) \right)_{i=1}^{d_j}, \tag{1.11}$$

where $\left(T_{d_{j-1}}^{w^{(j)}} \right)_{i,\cdot}$ denotes the ith row of the matrix $T_{d_{j-1}}^{w^{(j)}}$, $\left(b^{(j-1)} \right)_{\cdot,i}$ the ith column of the matrix $b^{(j-1)}$, and σ acts componentwisely on the components of the vector $\Phi^{(j-1)}(x) - \left(b^{(j-1)} \right)_{\cdot,i}$. That is, the ith component of $\Phi^{(j)}(x)$ equals the dot product of the ith row of $T_{d_{j-1}}^{w^{(j)}}$ and the action of σ on the difference of $\Phi^{(j-1)}(x)$ and the ith column of $b^{(j-1)}$.

Consider the special case when each row of the bias matrix (1.10) has identical components: $b_{i,i}^{(j-1)} = b_{i,1+i}^{(j-1)} = \ldots = b_{i,s+i}^{(j-1)} =: b_i^{(j-1)}$ for $i = 1, 2, \ldots, d_{j-1}$. Denote a vector $\vec{b}^{(j-1)} = \left(b_i^{(j-1)} \right)_{i=1}^{d_{j-1}}$. Then we see that in this special case, the vectors $\Phi^{(j-1)}(x) - \left(b^{(j-1)} \right)_{\cdot,i}$ and $\Phi^{(j-1)}(x) - \vec{b}^{(j-1)}$ are identical in their kth components with the indices k satisfying $\max\{1, i-s\} \le k \le \min\{d_{j-1}, i\}$. But the row vector $\left(T_{d_{j-1}}^{w^{(j)}} \right)_{i,\cdot}$ may have nonzero components only on this index set. Therefore, the equivalent form (1.11) of the iterative relationship (1.8) can be expressed as

$$\Phi^{(j)}(x) = \left(\left(T_{d_{j-1}}^{w^{(j)}} \right)_{i,\cdot} \sigma \left(\Phi^{(j-1)}(x) - \vec{b}^{(j-1)} \right) \right)_{i=1}^{d_j} = T_{d_{j-1}}^{w^{(j)}} \sigma \left(\Phi^{(j-1)}(x) - \vec{b}^{(j-1)} \right).$$

This is exactly the iterative relationship (1.2) for the DCNN stated in terms of the input vectors $\left([w]_{i,j} \cdot h^{(j-1)}(x) \right)_{i=1}^{d_{j-1}} = \Phi^{(j-1)}(x)$ with the weights given by $[w]_{i,j} = \left(w_{i-k}^{(j)} \right)_{k=1}^{d_{j-1}}$ in (1.3). Thus we see that the DCNN is a special case of the DDCNN. Moreover, the number of free parameters corresponding to the biases at level j is raised to $(s+1)d_{j-1}$ (for DDCNN) from d_{j-1} (for DCNN), having the same order. The total number of free parameters to be computed for the DDCNN is

$$(s+1)J + (s+1)\sum_{j=1}^{J-1} d_j + 2d_J = J\left(d + 1 + \frac{J+5}{2}s \right) + d + s\left(\frac{J(J-1)s}{2} + (J-1)d \right)$$

which is at most $s+1$ multiple of the number (1.5) for the DCNN. This justifies that the DDCNN has the same order of computational complexity as the DCNN.

The main analysis of this paper is to prove the following universality of approximation for the DDCNN, to be done in Section 5. Observe from the definition that $\mathcal{H}_J^{\mathbf{w},\mathbf{b}}$ consists of continuous piecewise linear functions (linear splines) on any compact subset Ω of $\mathbb{R}^d$. So the hypothesis space $\mathcal{H}_J^{\mathbf{w},\mathbf{b}}$ is a subset of $C(\Omega)$, the space of continuous functions on Ω.

Theorem 1. *Let $2 \le s \le d$. For any compact subset Ω of $\mathbb{R}^d$ and $f \in C(\Omega)$, there exist a sequence of filter masks $\mathbf{w}$ and a sequence of bias matrices $\mathbf{b}$ such that*

$$\lim_{J\to\infty} \inf_{f^* \in \mathcal{H}_J^{\mathbf{w},\mathbf{b}}} \left\{ \|f - f^*\|_{C(\Omega)} \right\} = 0. \tag{1.12}$$

2 Novelty of Analysis

To prove the universality of approximation stated in our main result, Theorem 1, by constructing a DDCNN, we first approximate the function f by a polynomial $P_\Gamma \in \mathcal{P}_\Gamma(\mathbb{R}^d)$ with some $\Gamma \in \mathbb{N}$ where $\mathcal{P}_\Gamma(\mathbb{R}^d)$ denotes the space of all polynomials on $\mathbb{R}^d$ of degree at most Γ.

Our first novelty is to observe that the polynomial P_Γ can be decomposed as

$$P_\Gamma(x) = \sum_{k=1}^{n_\Gamma} p_{k,\Gamma}(\xi_k \cdot x), \qquad x \in \mathbb{R}^d, \tag{2.1}$$

where each $p_{k,\Gamma} \in \mathcal{P}_\Gamma(\mathbb{R})$ is a univariate polynomial of degree at most Γ and $\{\xi_k\}_{k=1}^{n_\Gamma} \subset \mathbb{R}^d$ is a set of vectors, called **features**, with the number n_Γ depending only on d and Γ.

Our second novelty is to approximate the univariate polynomials $p_{k,\Gamma}$ by continuous piecewise linear functions (splines) spanned by $\{\sigma(\cdot - t_i)\}_{i=1}^N$ generated by a set of knots $\{t_1, \ldots, t_N\}$ to be chosen according to the approximation accuracy.

Our last and the most important novelty is to construct the filter masks $\mathbf{w} = \{w^{(j)}\}_{j=1}^J$ supported on $\{0, \ldots, s\}$ and the bias matrices $\mathbf{b} = (b^{(j)})_{j=1}^J$ in such a way that when the depth J is large enough, each polynomial $p_{k,\Gamma}(\xi_k \cdot x)$ in (2.1) can be approximated by functions from the hypothesis space $\mathcal{H}_J^{\mathbf{w},\mathbf{b}}$ of the DDCNN defined by (1.9).

Our construction of the DDCNN hypothesis space consists of three steps to be presented in the following three sections:

1. Construct the bias matrices $\mathbf{b} = (b^{(j)})_{j=1}^I$ and the initializing layers $\{\Phi^{(j)}(x)\}_{j=1}^I$ for a given set of filter masks $\{w^{(j)}\}_{j=1}^I$ so that the components of the last initializing layer $\Phi^{(I)}(x)$ can contain a rich family of linear functions $\{\eta_i \cdot x + \tau_i\}_{i=1}^{Is}$ with features $\{\eta_i\}_{i=1}^{Is} \subset \mathbb{R}^d$.

2. Construct deeper layers $\{\Phi^{(j)}(x)\}_{j=I+1}^{J}$ by a set of filter masks $\{w^{(j)}\}_{j=I+1}^{J}$ so that the components of the last deeper layer $\Phi^{(J)}(x)$ can contain the (spline) functions $\{\sigma(\eta_i \cdot x - t_j) : i = 1, \ldots, Is, j = 1, \ldots, N\}$ with an arbitrarily given knot sequence $t_1 < \ldots < t_N$.

3. Construct the filter masks $\{w^{(j)}\}_{j=1}^{I}$ in such a way that the set of features $\{\eta_i\}_{i=1}^{Is}$ constructed in the first step includes the set of features $\{\xi_k\}_{k=1}^{n_\Gamma}$ required in (2.1), and achieves a required approximation accuracy by constructing a spline approximation scheme to approximate the functions $p_{k,\Gamma}(\xi_k \cdot x)$ by linear combinations of $\{\sigma(\eta_i \cdot x - t_j)\}$ from the hypothesis space.

Throughout our construction, we assume that $s \le d$ and the filter mask sequence $\{w^{(j)}\}$ satisfies $w_0^{(j)} > 0$ for each j and $w_s^{(j)} > 0$ for each $j = I+1, \ldots, J$, which will be realized in our mask factorization in Section 5. Some ideas of our construction are from ridge approximation, wavelets, spline functions, and learning theory [8, 24].

In practical applications of DCNNs, fully connected layers might be added and techniques like polling are involved [9]. These technical tools for empirical implementations will not be discussed in this paper.

3 Initializing Layers

To give a linear algebra viewpoint of our construction, we use the matrices $T_{d_{j-1}}^{w^{(j)}}$ defined by (1.4). From the definition (1.7), we see that the first input vector can be written as

$$\Phi^{(1)}(x) = T_d^{w^{(1)}} x. \tag{3.1}$$

From the assumption $w_0^{(1)} \neq 0$, we find the linear independence of the first d rows of the matrix $T_d^{w^{(1)}}$. Hence the first d components of $\Phi^{(1)}$, $\{\phi_i^{(1)}(x) = w_0^{(1)} x_i + \sum_{k=\max\{1,i-s\}}^{i-1} w_{i-k}^{(1)} x_k\}_{i=1}^{d}$, span the other components and the space of all homogeneous linear polynomials on Ω. Observe that each component $\phi_i^{(1)}(x)$ of the first layer contains at most $s+1$ variables. This special property motivates us to construct $\{\Phi^{(j)}\}_{j=1}^{I}$, the first I layers of input vectors with $I \in \mathbb{N}$, called **initializing layers**, in such a way that the components of the last initialing layer $\Phi^{(I)}$ are linear functions involving more variables with some coefficient vectors being the features required for (2.1).

Extending the matrix form (3.1) of $\Phi^{(1)}$, we define **homogenized layers** as

$$\widehat{\Phi}^{(j)}(x) = T_{d_{j-1}}^{w^{(j)}} \ldots T_d^{w^{(1)}} x, \qquad j \in \mathbb{N}. \tag{3.2}$$

We construct the initializing layers by taking the biases $\{b_{k,i}^{(j-1)}\}$ in (1.8) to be small enough so that $\phi_k^{(j-1)}(x) - b_{k,i}^{(j-1)} \ge 0$. Denote the ℓ^1-norm of a finitely supported

sequence w on $\mathbb{Z}$ as $\|w\|_1 = \sum_{k\in\mathbb{Z}} |w_k|$. Denote

$$B^{(0)} = \max_{x\in\Omega} \max_{k=1,\dots,d} |x_k|, \qquad B^{(j)} = \|w^{(j)}\|_1 \dots \|w^{(1)}\|_1 B^{(0)}, \qquad j \in \mathbb{N}.$$

Theorem 2. *For $j \in \mathbb{N}$, we have*

$$\left\| \left(\widehat{\Phi}^{(j)}(x) \right)_i \right\|_{C(\Omega)} \le B^{(j)}, \qquad \forall i = 1, \dots, d_j. \tag{3.3}$$

Let $I \in \mathbb{N}$. Set vectors $\tau^{(j)} \in \mathbb{R}^{d_j}$ by

$$\tau_i^{(j)} = \sum_{k=\max\{1,i-s\}}^{\min\{d_{j-1},i\}} w_{i-k}^{(j)} B^{(j-1)}, \qquad i = 1, \dots, d_j,\ j = 2, \dots, I.$$

Then all the components of $\tau^{(j)}$ lie in $[-B^{(j)}, B^{(j)}]$. Construct the bias matrices $\{b^{(j)}\}_{j=1}^{I-1}$ by

$$b_{k,i}^{(j)} = \tau_k^{(j)} - B^{(j)}, \qquad \forall\ k = \max\{1, i-s\}, \dots, \min\{d_j, i\},\ i = 1, \dots, d_j, \tag{3.4}$$

where $\tau^{(1)}$ denotes the zero vector. Then we have

$$\Phi^{(j)}(x) = T_{d_{j-1}}^{w^{(j)}} \dots T_d^{w^{(1)}} x + \tau^{(j)} = \widehat{\Phi}^{(j)}(x) + \tau^{(j)}, \qquad \forall\ j = 1, \dots, I. \tag{3.5}$$

Proof. From the definition of $\widehat{\Phi}^{(1)} = \Phi^{(1)}$, we know that

$$\left\| \left(\widehat{\Phi}^{(1)}(x) \right)_i \right\|_{C(\Omega)} = \|\widehat{\phi}_i^{(1)}\|_{C(\Omega)} \le \|w^{(1)}\|_1 B^{(0)} = B^{(1)}, \qquad \forall i = 1, \dots, d_1.$$

The definition (3.2) yields the iteration relation $\widehat{\Phi}^{(j+1)}(x) = T_{d_j}^{w^{(j+1)}} \widehat{\Phi}^{(j)}(x)$. Hence

$$\left| \widehat{\phi}_i^{(j+1)}(x) \right| = \left| \sum_{k=1}^{d_j} w_{i-k}^{(j+1)} \widehat{\phi}_k^{(j)}(x) \right| \le \|w^{(j+1)}\|_1 \max_{k=1,\dots,d_j} \left\| \left(\widehat{\Phi}^{(j)}(x) \right)_k \right\|_{C(\Omega)}.$$

This proves (3.3) by induction on j.

From the definition of the norm, we know that $\left| \tau_i^{(j)} \right| \le \|w^{(j)}\|_1 B^{(j-1)} = B^{(j)}$. Hence all the components of $\tau^{(j)}$ lie in $[-B^{(j)}, B^{(j)}]$.

To prove (3.5) with $j = 2$ for the second layer, we find from the choice (3.4) of the biases and the notation $\tau^{(1)} = 0$ that $b_{k,i}^{(1)} = -B^{(1)}$ in (1.8), and $\phi_k^{(1)}(x) - b_{k,i}^{(1)} = \phi_k^{(1)}(x) + B^{(1)} \ge 0$ for $i = 1, \dots, d_2$, which implies

$$\phi_i^{(2)}(x) = \sum_{k=\max\{1,i-s\}}^{\min\{d_1,i\}} w_{i-k}^{(2)} \phi_k^{(1)}(x) + \sum_{k=\max\{1,i-s\}}^{\min\{d_1,i\}} w_{i-k}^{(2)} B^{(1)}.$$

Due to the mask support property (1.6) and (3.1), we know that

$$\sum_{k=\max\{1,i-s\}}^{\min\{d_1,i\}} w_{i-k}^{(2)}\phi_k^{(1)}(x) = \sum_{k=1}^{d_1} w_{i-k}^{(2)}\left(T_d^{w^{(1)}}x\right)_k = \left(T_{d_1}^{w^{(2)}}T_d^{w^{(1)}}x\right)_i.$$

Thus (3.5) holds true for $j=2$.

Now we prove (3.5) by induction. Assume that $\Phi^{(j)}(x) = T_{d_{j-1}}^{w^{(j)}}\dots T_d^{w^{(1)}}x + \tau^{(j)}$. From this induction hypothesis and the choice (3.4) of the biases for (1.8), we find $\phi_k^{(j)}(x) - b_{k,i}^{(j)} = \widehat{\phi}_k^{(j)}(x) + B^{(j)} \geq 0$ by (3.3). It follows from (1.6) again that for $i = 1, \dots, d_{j+1}$,

$$\begin{aligned}\phi_i^{(j+1)}(x) &= \sum_{k=\max\{1,i-s\}}^{\min\{d_j,i\}} w_{i-k}^{(j+1)}\widehat{\phi}_k^{(j)}(x) + \sum_{k=\max\{1,i-s\}}^{\min\{d_j,i\}} w_{i-k}^{(j+1)}B^{(j)} \\ &= \sum_{k=1}^{d_j} w_{i-k}^{(j+1)}\left(T_{d_{j-1}}^{w^{(j)}}\dots T_d^{w^{(1)}}x\right)_k + \tau_i^{(j+1)} = \left(T_{d_j}^{w^{(j+1)}}\dots T_d^{w^{(1)}}x\right)_i + \tau_i^{(j+1)}.\end{aligned}$$

This completes the induction procedure and proves our conclusion. □

The last initializing layer can be expressed as

$$\Phi^{(I)}(x) = \widehat{\Phi}^{(I)}(x) + \tau^{(I)} = T_{d_{I-1}}^{w^{(I)}}\dots T_d^{w^{(1)}}x + \tau^{(I)}, \tag{3.6}$$

where $T_{d_{I-1}}^{w^{(I)}}\dots T_d^{w^{(1)}}$ is a $d_I \times d$ matrix. This matrix can be written in the form (1.4). To see this, we recall the convolution $a*b$ of two finitely supported sequences a, b on $\mathbb{Z}$ defined by

$$(a*b)_i = \sum_{k\in\mathbb{Z}} a_{i-k}b_k, \qquad i \in \mathbb{Z}.$$

If a is supported on $\{A_0, A_0+1, \dots, A_1\}$ and b on $\{B_0, B_0+1, \dots, B_1\}$, then $a*b$ is supported on $\{A_0+B_0, A_0+B_0+1, \dots, A_1+B_1\}$. Now we consider the matrix product $T_{d_{j-1}}^{w^{(j)}}T_{d_{j-2}}^{w^{(j-1)}}$ which is a $d_j \times d_{j-2}$ matrix. Its (i,k)-entry with $1 \leq i \leq d_j$ and $1 \leq k \leq d_{j-2}$ can be expressed as

$$\left(T_{d_{j-1}}^{w^{(j)}}T_{d_{j-2}}^{w^{(j-1)}}\right)_{i,k} = \sum_{\ell=1}^{d_{j-1}}\left(T_{d_{j-1}}^{w^{(j)}}\right)_{i,\ell}\left(T_{d_{j-2}}^{w^{(j-1)}}\right)_{\ell,k} = \sum_{\ell=1}^{d_{j-1}} w_{i-\ell}^{(j)}w_{\ell-k}^{(j-1)}.$$

For $\ell \leq 0$, we have $\ell - k < 0$ and thereby $w_{\ell-k}^{(j-1)} = 0$, while for $\ell \geq d_{j-1}+1$, we have $\ell - k > s$ which also implies $w_{\ell-k}^{(j-1)} = 0$. Therefore,

$$\left(T_{d_{j-1}}^{w^{(j)}}T_{d_{j-2}}^{w^{(j-1)}}\right)_{i,k} = \sum_{\ell\in\mathbb{Z}} w_{i-\ell}^{(j)}w_{\ell-k}^{(j-1)} = \sum_{p\in\mathbb{Z}} w_{i-k-p}^{(j)}w_p^{(j-1)} = \left(w^{(j)}*w^{(j-1)}\right)_{i-k}.$$

This is exactly the (i,k)-entry of the $d_j \times d_{j-2}$ matrix $T_{d_{j-2}}^{w^{(j)}*w^{(j-1)}}$ defined by (1.4) corresponding to the filter mask $w^{(j)}*w^{(j-1)}$ supported on $\{0,\ldots,2s\}$. Hence $T_{d_{j-1}}^{w^{(j)}} T_{d_{j-2}}^{w^{(j-1)}} = T_{d_{j-2}}^{w^{(j)}*w^{(j-1)}}$. Applying this relation iteratively tells us that

$$T_{d_{I-1}}^{w^{(I)}} \ldots T_d^{w^{(1)}} = T_d^W \in \mathbb{R}^{(d+Is)\times d},$$

a matrix of the form (1.4) with $D = d$ and s replaced by Is, where $W = W^{(I)} : \mathbb{Z} \to \mathbb{R}$ is a sequence supported on $\{0,\ldots,Is\}$ defined by

$$W = w^{(I)}*w^{(I-1)}*\ldots*w^{(2)}*w^{(1)}. \tag{3.7}$$

Note that $W_0 = \Pi_{j=1}^I w_0^{(j)} > 0$. At the end, when $I \geq \frac{d}{s}$, the I-th homogenized layer corresponding to the last initializing layer is given by $\widehat{\Phi}^{(I)}(x) = T_d^W x$ or more explicitly by

$$T_d^W = \begin{bmatrix} W_0 & 0 & \cdots & 0 \\ W_1 & W_0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ W_{d-1} & \cdots & W_1 & W_0 \\ W_d & W_{d-1} & \cdots & W_1 \\ \vdots & \ddots & \ddots & \vdots \\ W_{Is} & \cdots & \ddots & \vdots \\ 0 & W_{Is} & \cdots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & W_{Is} \end{bmatrix}, \; \widehat{\Phi}^{(I)} = \begin{bmatrix} \widehat{\phi}_1^{(I)}(x) = W_0 x_1 \\ \vdots \\ \widehat{\phi}_{d-1}^{(I)}(x) = W_{d-2}x_1 + \ldots + W_0 x_{d-1} \\ \widehat{\phi}_d^{(I)}(x) = W_{d-1}x_1 + \ldots + W_0 x_d \\ \widehat{\phi}_{d+1}^{(I)}(x) = W_d x_1 + \ldots + W_1 x_d \\ \vdots \\ \widehat{\phi}_{d+\ell}^{(I)}(x) = W_{d+\ell-1}x_1 + \ldots + W_\ell x_d \\ \vdots \\ \widehat{\phi}_{Is+1}^{(I)}(x) = W_{Is}x_1 + \ldots + W_{Is-d+1}x_d \\ \vdots \\ \widehat{\phi}_{d+Is}^{(I)}(x) = W_{Is}x_d \end{bmatrix}. \tag{3.8}$$

Hence the last Is components of the last homogenized layer $\widehat{\Phi}^{(I)}(x)$ can be expressed as

$$\{\widehat{\Phi}_{d+i}^{(I)}(x)\}_{i=1}^{Is} = \{\eta_i \cdot x\}_{i=1}^{Is}, \quad \text{with } \eta_i = (W_{d-1+i}, W_{d-2+i}, \ldots, W_i)^T \in \mathbb{R}^d.$$

Note the first $i-1$ components of η_i vanish for $i > Is-d$. In our proof of universality of approximation, the set of features $\{\eta_{(k-1)d+1}\}_{1\leq k\leq (Is-1)/d} \subset \mathbb{R}^d$ will be used.

4 Deeper Layers for Constructing Linear Splines

Let $t_1 < t_2 < \ldots < t_{N-1} < t_N$. Based on the first I (initializing) layers, we shall construct the next $(N-1)I$ layers, called **deeper layers**, of input vectors of a DDCNN of depth $J = NI$ so that functions of type $f(\phi_{d+\ell}^{(I)}(x))$ with $\ell \in \{1,\ldots,Is\}$ and f

being a continuous linear spline supported on $[t_1, t_N]$ with knots $\{t_1, t_2, \ldots, t_{N-1}, t_N\}$ are contained in the hypothesis space $\mathcal{H}_J^{\mathbf{w},\mathbf{b}}$.

We take a set of filter masks $\{w^{(j)}\}_{j=I+1}^{NI}$ supported on $\{0, \ldots, s\}$ to satisfy $w_0^{(j)} > 0$ and $w_s^{(j)} > 0$ for $j = I+1, \ldots, NI$. The constructed DDCNN input vectors $\{\Phi^{(j)}(x)\}_{j=I+1}^{NI}$ of deeper layers will be expressed in block forms in terms of the following two types of blocks:

$$\mathcal{L} := \begin{bmatrix} \widehat{\phi}_1^{(I)}(x) + B^{(I)} \\ \vdots \\ \widehat{\phi}_d^{(I)}(x) + B^{(I)} \end{bmatrix} = \begin{bmatrix} W_0 x_1 + B^{(I)} \\ \vdots \\ W_{d-1} x_1 + \ldots + W_0 x_d + B^{(I)} \end{bmatrix} \tag{4.1}$$

and

$$\Sigma_{\ell,t} := \left[\sigma(\widehat{\phi}_i^{(I)}(x) - t)\right]_{i=d_\ell - s+1}^{d_\ell} = \begin{bmatrix} \sigma\left(\widehat{\phi}_{d+(\ell-1)s+1}^{(I)}(x) - t\right) \\ \vdots \\ \sigma\left(\widehat{\phi}_{d+\ell s}^{(I)}(x) - t\right) \end{bmatrix}, \qquad \ell \in \{1, \ldots, I\}. \tag{4.2}$$

To illustrate our construction explicitly, we express the $d_j \times d_{j-1}$ matrix $T_{d_{j-1}}^{w^{(j)}}$ defined by (1.4) in the following block matrix form

$$T_{d_{j-1}}^{w^{(j)}} = \begin{bmatrix} L_0 & O & \cdots & \cdots & \cdots & \cdots & \cdots & O \\ O & U_1 & L_1 & O & \cdots & \cdots & \cdots & O \\ O & & U_2 & L_2 & O & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & U_q & L_q & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & U_{q+1} & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots & O \\ \vdots & & \ddots & \ddots & \ddots & \cdots & U_{j-1} & L_{j-1} \\ O & & \cdots & \cdots & O & \cdots & O & U_j \end{bmatrix}, \tag{4.3}$$

where O denotes a zero matrix which might have different sizes in various occurrences, $L_0 = \left[w_{\ell-m}^{(j)}\right]_{\ell,m=1}^{d}$ is a lower triangular $d \times d$ matrix given by

$$L_0 = \begin{bmatrix} w_0^{(j)} & 0 & \cdots & \cdots & & 0 \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ w_s^{(j)} & \cdots & w_0^{(j)} & 0 \cdots & & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ 0 & \cdots & 0 & w_s^{(j)} & \cdots & w_0^{(j)} \end{bmatrix}_{d \times d}, \tag{4.4}$$

$L_1 = \ldots = L_{j-1} = L$ are identical lower triangular $s \times s$ matrices and $U_1 = \ldots = U_j = U$ are identical upper triangular $s \times s$ matrices given by

$$L = \left[w_{\ell-m}^{(j)}\right]_{\ell,m=1}^{s} = \begin{bmatrix} w_0^{(j)} & 0 & \cdots \\ \vdots & \ddots & \ddots \\ w_{s-1}^{(j)} & \cdots & w_0^{(j)} \end{bmatrix}, \; U = \left[w_{s+\ell-m}^{(j)}\right]_{\ell,m=1}^{s} = \begin{bmatrix} w_s^{(j)} & \cdots & w_1^{(j)} \\ \vdots & \ddots & \ddots \\ \cdots & 0 & w_s^{(j)} \end{bmatrix}. \tag{4.5}$$

In our construction described in details in the following subsections, we take suitable biases in (1.8) so that only one term in the summation for $\phi_i^{(j)}(x)$ does not vanish.

4.1 First deeper layer

In our construction of the first deeper layer $\Phi^{(I+1)}(x)$, for each component $\phi_i^{(I+1)}(x)$ expressed in (1.8), we take all biases except one to be $b_{k,i}^{(I)} = 2B^{(I)}$ which together with Theorem 2 implies $\phi_k^{(I)}(x) - b_{k,i}^{(I)} \leq 0$ and $w_{i-k}^{(I+1)} \sigma\left(\phi_k^{(I)}(x) - b_{k,i}^{(I)}\right) = 0$. In a linear algebra viewpoint taking the matrix $T_{d_I}^{w^{(I+1)}}$ in the block form (4.3) with $q = 1$, the exceptional bias for each component $\phi_i^{(I+1)}(x)$ is taken corresponding to the main diagonal entry $w_0^{(I+1)}$ of L_0 (for $1 \leq i \leq d$) or L_1 (for $d+1 \leq i \leq d+s$), or the main diagonal entry $w_s^{(I+1)}$ of the block U_ℓ (for $d + (\ell-1)s + 1 \leq i \leq d + \ell s$) with $\ell \in \{2, \ldots, I+1\}$. Here the blocks L_1 and U_2 lie on the same column group, so the same block $\left[\widehat{\phi}_i^{(I)}(x)\right]_{i=d+1}^{d+s}$ of components of $\Phi^{(I)}(x)$ is used to generate two blocks of $\Phi^{(I+1)}(x)$ involving Σ_{1,t_1} and Σ_{1,t_2}. The exact expression for the first deeper layer $\Phi^{(I+1)}(x)$ in terms of the block matrices $\mathcal{L}$ and $\Sigma_{\ell,t}$ is the following.

Lemma 1. *By taking the bias matrix $b^{(I)}$ as*

$$b_{k,i}^{(I)} = \begin{cases} \tau_i^{(I)} - B^{(I)}, & \text{if } 1 \leq i \leq d \text{ and } k = i, \\ \tau_i^{(I)} + t_1, & \text{if } d+1 \leq i \leq d+s \text{ and } k = i, \\ \tau_{i-s}^{(I)} + t_2, & \text{if } d+s+1 \leq i \leq d+2s \text{ and } k = i-s, \\ \tau_{i-s}^{(I)} + t_1, & \text{if } d+2s+1 \leq i \leq d+(I+1)s \text{ and } k = i-s, \\ 2B^{(I)}, & \text{otherwise}, \end{cases}$$

we have

$$\Phi^{(I+1)}(x) = \begin{bmatrix} w_0^{(I+1)} \mathcal{L} \\ w_0^{(I+1)} \Sigma_{1,t_1} \\ w_s^{(I+1)} \Sigma_{1,t_2} \\ \left[w_s^{(I+1)} \Sigma_{\ell,t_1}\right]_{\ell=2}^{I} \end{bmatrix}. \tag{4.6}$$

Proof. With the given choice of the bias matrix $b^{(I)}$, we find that in the summation for each component $\phi_i^{(I+1)}(x)$ of $\Phi^{(I+1)}(x)$ defined in (1.8), only one term does not vanish and this term corresponds to $k = i$ or $k = i - s$, which yields

$$\phi_i^{(I+1)}(x) = \begin{cases} w_0^{(I+1)} \left(\widehat{\phi}_i^{(I)}(x) + B^{(I)} \right), & \text{if } 1 \le i \le d, \\ w_0^{(I+1)} \sigma \left(\widehat{\phi}_i^{(I)}(x) - t_1 \right), & \text{if } d+1 \le i \le d+s, \\ w_s^{(I+1)} \sigma \left(\widehat{\phi}_{i-s}^{(I)}(x) - t_2 \right), & \text{if } d+s+1 \le i \le d+2s, \\ w_s^{(I+1)} \sigma \left(\widehat{\phi}_{i-s}^{(I)}(x) - t_1 \right), & \text{if } d+2s+1 \le i \le d+(I+1)s. \end{cases}$$

This together with the definition of the block matrices $\mathcal{L}$ and $\Sigma_{\ell,t}$ verifies the stated expression for $\Phi^{(I+1)}(x)$. □

Notice that for the first deeper layer $\widehat{\Phi}^{(I+1)}(x)$, the first d components are linear functions and the remaining $(I+1)s$ components are splines with knots t_1 or t_2 composed with the corresponding components of $\widehat{\Phi}^{(I)}(x)$.

4.2 Constructing deeper layers and splines with two knots

For constructing more deeper layers $\{\Phi^{(I+\ell)}(x)\}_{\ell=2}^{I}$, we observe from the expression (4.6) for $\Phi^{(I+1)}(x)$ that the block $\left[\widehat{\phi}_i^{(I+1)}(x)\right]_{i=d_2+1}^{d_3}$ involving Σ_{2,t_1} should be used to generate two blocks involving Σ_{2,t_1} and Σ_{2,t_2}. This motivates our idea of considering the block $\left[\phi_i^{(I+\ell-1)}(x)\right]_{i=d_{2(\ell-1)}+1}^{d_{2\ell-1}}$ of $\Phi^{(I+\ell-1)}(x)$ for constructing $\Phi^{(I+\ell)}(x)$, and emphasizing the corresponding blocks $L_{2\ell-1}$ and $U_{2\ell}$ in the block form (4.3) of the matrix $T_{d_{I+\ell-1}}^{w^{(I+\ell)}}$ for taking exceptional biases.

We need the following properties for the ReLU σ:

$$\sigma(\alpha u) = \alpha\sigma(u), \qquad \sigma(\sigma(u) - \alpha) = \sigma(u - \alpha), \qquad \forall\ u \in \mathbb{R}, \alpha > 0. \tag{4.7}$$

Denote

$$[\Pi_0]_i^j = \Pi_{p=i}^j w_0^{(p)}, \qquad [\Pi_s]_i^j = \Pi_{p=i}^j w_s^{(p)}.$$

Observe that for $I+1 \le i \le j$, both $[\Pi_0]_i^j$ and $[\Pi_s]_i^j$ are positive by our assumptions on $w_0^{(j)}$ and $w_s^{(j)}$.

Lemma 2. *There exist bias matrices $\left\{b^{(j)}\right\}_{j=I}^{2I-1}$ (constructed explicitly) such that*

$$\Phi^{(I+\ell)}(x) = \begin{bmatrix} [\Pi_0]_{I+1}^{I+\ell}\mathcal{L} \\ \begin{bmatrix} [\Pi_s]_{I+1}^{I+j-1}[\Pi_0]_{I+j}^{I+\ell} & \Sigma_{j,t_1} \\ [\Pi_s]_{I+1}^{I+j}[\Pi_0]_{I+j+1}^{I+\ell} & \Sigma_{j,t_2} \end{bmatrix}_{j=1}^{\ell} \\ \left[[\Pi_s]_{I+1}^{I+\ell} \ \Sigma_{j,t_1} \right]_{j=\ell+1}^{I} \end{bmatrix}, \qquad \ell = 1, \ldots, I. \tag{4.8}$$

Proof. We prove our statement by induction. The case $\ell = 1$ has been shown in Lemma 1.

Assume that the expression (4.8) holds for $\ell - 1$. Then the first components of $\Phi^{(I+\ell-1)}(x)$ are bounded by $[\Pi_0]_{I+1}^{I+\ell-1} 2B^{(I)} \leq 2\|w^{(I+\ell-1)}\|_1 \ldots \|w^{(I+1)}\|_1 B^{(I)} = 2B^{(I+\ell-1)}$, while the other components are bounded by

$$\Pi_{p=I+1}^{I+\ell-1}\|w^{(p)}\|_1 \left(\max_i \|\widehat{\phi}_i^{(I)}\|_{C(\Omega)} + \max\{|t_1|, |t_2|\}\right) \leq B^{(I+\ell-1)} \left(2 + \frac{\max\{|t_1|, |t_2|\}}{B^{(I)}}\right).$$

Denote this upper bound by $D_{I+\ell-1}$.

To verify the statement for the case ℓ, by viewing the blocks $L_{2\ell-1}$ and $U_{2\ell}$ of the matrix form (4.3) with $q = 2\ell - 1$ for $T_{d_{I+\ell-1}}^{w^{(I+\ell)}}$, we take the entries of the bias matrix $b^{(I+\ell-1)}$ in (1.8) with special entries corresponding to $k = i$ or $k = i - s$ as

$$b_{k,i}^{(I+\ell-1)} = \begin{cases} 0, & \text{if } 1 \leq i \leq d_{2\ell-1} \text{ and } k = i, \\ [\Pi_s]_{I+1}^{I+\ell-1}(t_2 - t_1), & \text{if } d_{2\ell-1} + 1 \leq i \leq d_{2\ell} \text{ and } k = i - s, \\ 0, & \text{if } d_{2\ell} + 1 \leq i \leq d_{I+\ell} \text{ and } k = i - s, \\ D_{I+\ell-1}, & \text{otherwise.} \end{cases}$$

It is essential to note that all the components of $\Phi^{(I+\ell-1)}(x)$ are nonnegative. So we see from property (4.7) of the ReLU σ that the first $d_{2\ell-1}$ components of $\Phi^{(I+\ell)}(x)$ are just the $w_0^{(I+\ell)}$ multiples of those of $\Phi^{(I+\ell-1)}(x)$, and the last $(I - \ell)s$ components of $\Phi^{(I+\ell)}(x)$ are the $w_s^{(I+\ell)}$ multiples of those of $\Phi^{(I+\ell-1)}(x)$. For the middle s components, we see from the definition (1.8) and the choice of the bias matrix that

$$\phi_i^{(I+\ell)}(x) = w_s^{(I+\ell)}\sigma\left(\phi_{i-s}^{(I+\ell-1)}(x) - [\Pi_s]_{I+1}^{I+\ell-1}(t_2 - t_1)\right), \qquad d_{2\ell-1} + 1 \leq i \leq d_{2\ell}.$$

By the induction hypothesis, $\left[\phi_{i-s}^{(I+\ell-1)}(x)\right]_{i=d_{2\ell-1}+1}^{d_{2\ell}}$ is the block matrix $[\Pi_s]_{I+1}^{I+\ell-1}\Sigma_{\ell,t_1}$. So for $\alpha = 1, \ldots, s$, we have $\phi_{d_{2\ell-1}-s+\alpha}^{(I+\ell-1)}(x) = [\Pi_s]_{I+1}^{I+\ell-1}\sigma\left(\widehat{\phi}_{d_\ell-s+\alpha}^{(I)}(x) - t_1\right)$. Thus by (4.7) there holds

$$\begin{aligned} \phi_{d_{2\ell-1}+\alpha}^{(I+\ell)}(x) &= w_s^{(I+\ell)}\sigma\left([\Pi_s]_{I+1}^{I+\ell-1}\sigma\left(\widehat{\phi}_{d_\ell-s+\alpha}^{(I)}(x) - t_1\right) - [\Pi_s]_{I+1}^{I+\ell-1}(t_2 - t_1)\right) \\ &= w_s^{(I+\ell)}[\Pi_s]_{I+1}^{I+\ell-1}\sigma\left(\sigma\left(\widehat{\phi}_{d_\ell-s+\alpha}^{(I)}(x) - t_1\right) - (t_2 - t_1)\right) \\ &= [\Pi_s]_{I+1}^{I+\ell}\sigma\left(\widehat{\phi}_{d_\ell-s+\alpha}^{(I)}(x) - t_2\right), \qquad \alpha = 1, \ldots, s. \end{aligned}$$

Hence $\left[\phi_i^{(I+\ell)}(x)\right]_{i=d_{2\ell-1}+1}^{d_{2\ell}}$ equals the block matrix $[\Pi_s]_{I+1}^{I+\ell}\Sigma_{\ell,t_2}$. Therefore the desired expression (4.8) holds true for the case ℓ. This completes the induction procedure. □

Lemma 2 yields the following expression for $\Phi^{(2I)}(x)$:

$$\Phi^{(2I)}(x) = \begin{bmatrix} [\Pi_0]_{I+1}^{2I}\mathcal{L} \\ \mathcal{S}_1^{(2)} \\ \vdots \\ \mathcal{S}_I^{(2)} \end{bmatrix}, \text{ where } \mathcal{S}_j^{(2)} = \begin{bmatrix} [\Pi_s]_{I+1}^{I+j-1}[\Pi_0]_{I+j}^{2I} & \Sigma_{j,t_1} \\ [\Pi_s]_{I+1}^{I+j}[\Pi_0]_{I+j+1}^{2I} & \Sigma_{j,t_2} \end{bmatrix}. \tag{4.9}$$

Note that for the block $\mathcal{S}_j^{(2)}$, the components of $\Sigma_{j,t_1} = \left[\sigma(\widehat{\phi}_i^{(I)}(x) - t_1)\right]_{i=d_j-s+1}^{d_j}$ and Σ_{j,t_2} are splines $\sigma(\cdot - t_1)$ with knots t_1 and $\sigma(\cdot - t_2)$ with knots t_2 respectively composed with the linear functions $\left\{\widehat{\phi}_i^{(I)}(x)\right\}_{i=d_j-s+1}^{d_j}$.

4.3 Constructing more deeper layers and splines with more knots

We now apply the above procedure to construct further deeper layers and create splines with knots $t_3, \ldots, t_N$. For $p \geq 2, 1 \leq k \leq p, 1 \leq j \leq I$, denote

$$\Pi_{p,j,k} = \Pi_{\nu=1}^{k-1}\left\{[\Pi_s]_{\nu I+1}^{\nu I+j}[\Pi_0]_{\nu I+j+1}^{(\nu+1)I}\right\}\Pi_{\nu=k}^{p-1}\left\{[\Pi_s]_{\nu I+1}^{\nu I+j-1}[\Pi_0]_{\nu I+j}^{(\nu+1)I}\right\}.$$

When $p = 2$, the numbers $\Pi_{2,j,1}$ and $\Pi_{2,j,2}$ are exactly those in (4.9). Observe that $0 < \Pi_{p,j,k} \leq \Pi_{\ell=I+1}^{pI}\|w^{(\ell)}\|_1 = \frac{B^{(pI)}}{B^{(I)}}$.

Lemma 3. *Let $2 \leq N \in \mathbb{N}$. There exist bias matrices $\left\{b^{(j)}\right\}_{j=2I}^{NI-1}$ (constructed explicitly) such that for $p = 2, \ldots, N$, there holds*

$$\Phi^{(pI)}(x) = \begin{bmatrix} [\Pi_0]_{I+1}^{pI}\mathcal{L} \\ \mathcal{S}_1^{(p)} \\ \vdots \\ \mathcal{S}_I^{(p)} \end{bmatrix}, \text{ where } \mathcal{S}_j^{(p)} = \begin{bmatrix} \Pi_{p,j,1}\Sigma_{j,t_1} \\ \vdots \\ \Pi_{p,j,p}\Sigma_{j,t_p} \end{bmatrix}. \tag{4.10}$$

Proof. We prove by induction for p. The case $p = 2$ has been shown in Lemma 2.

Suppose that the statement holds true for $p - 1$.

To prove the statement for the case p, we assume that for some $\ell \in \{1, \ldots, I\}$,

$$\Phi^{((p-1)I+\ell-1)}(x) = \begin{bmatrix} [\Pi_0]_{I+1}^{(p-1)I+\ell-1}\mathcal{L} \\ \mathcal{S}_1^{(p-1,\ell-1)} \\ \vdots \\ \mathcal{S}_I^{(p-1,\ell-1)} \end{bmatrix}, \tag{4.11}$$

which is true for $\ell = 1$ due to the induction hypothesis for $\Phi^{((p-1)I)}(x)$. Here

$$\mathcal{S}_j^{(p-1,\ell-1)} = \begin{cases} \begin{bmatrix} [\Pi_s]_{(p-1)I+1}^{(p-1)I+j-1} [\Pi_0]_{(p-1)I+j}^{(p-1)I+\ell-1} \Pi_{p-1,j,1}\Sigma_{j,t_1} \\ \vdots \\ [\Pi_s]_{(p-1)I+1}^{(p-1)I+j-1} [\Pi_0]_{(p-1)I+j}^{(p-1)I+\ell-1} \Pi_{p-1,j,p-1}\Sigma_{j,t_{p-1}} \\ [\Pi_s]_{(p-1)I+1}^{(p-1)I+j} [\Pi_0]_{(p-1)I+j+1}^{(p-1)I+\ell-1} \Pi_{p-1,j,p-1}\Sigma_{j,t_p} \end{bmatrix}, & \text{if } j = 1,\ldots,\ell-1, \\ \begin{bmatrix} [\Pi_s]_{(p-1)I+1}^{(p-1)I+\ell-1} \Pi_{p-1,j,1}\Sigma_{j,t_1} \\ \vdots \\ [\Pi_s]_{(p-1)I+1}^{(p-1)I+\ell-1} \Pi_{p-1,j,p-1}\Sigma_{j,t_{p-1}} \end{bmatrix}, & \text{if } j = \ell,\ldots,I. \end{cases}$$

Then by property (4.7) of the ReLU and Theorem 2, we have

$$0 \le \phi_k^{((p-1)I+\ell-1)}(x) \le \frac{B^{((p-1)I+\ell-1)}}{B^{(I)}} \max\left\{2B^{(I)}, B^{(I)} + \max\{|t_1|,\ldots,|t_p|\}\right\} =: D_{p,\ell-1}.$$

To construct $\Phi^{((p-1)I+\ell)}(x)$, we choose the bias matrix $b^{((p-1)I+\ell-1)}$ for (1.8) with special choices corresponding to the main diagonal entries of the blocks $L_{p\ell-1}$ and $U_{p\ell}$ of the matrix form (4.3) with $q = p\ell - 1$ for $T_{d_{(p-1)I+\ell-1}}^{w^{((p-1)I+\ell)}}$ by setting $b_{k,i}^{((p-1)I+\ell-1)}$ equal to

$$\begin{cases} 0, & \text{if } 1 \le i \le d_{p\ell-1} \text{ and } k = i, \\ [\Pi_s]_{(p-1)I+1}^{(p-1)I+\ell-1} \Pi_{p-1,\ell,p-1}\left(t_p - t_{p-1}\right), & \text{if } d_{p\ell-1}+1 \le i \le d_{p\ell} \text{ and } k = i-s, \\ 0, & \text{if } d_{p\ell}+1 \le i \le d_{(p-1)I+\ell} \text{ and } k = i-s, \\ D_{p,\ell-1}, & \text{otherwise.} \end{cases}$$

It follows from property (4.7) of the ReLU and $0 \le \phi_k^{((p-1)I+\ell-1)}(x) \le D_{p,\ell-1}$ that the first $d_{p\ell-1}$ components of $\Phi^{((p-1)I+\ell)}(x)$ are just the $w_0^{((p-1)I+\ell)}$ multiples of those of $\Phi^{((p-1)I+\ell-1)}(x)$, while the last $(p-1)(I-\ell)s$ components of $\Phi^{((p-1)I+\ell)}(x)$ are the $w_s^{((p-1)I+\ell)}$ multiples of those of $\Phi^{((p-1)I+\ell-1)}(x)$. For the middle s components with indices $d_{p\ell-1}+1 \le i \le d_{p\ell}$, we have

$$\phi_i^{((p-1)I+\ell)}(x) = w_s^{((p-1)I+\ell)}\sigma\left(\phi_{i-s}^{((p-1)I+\ell-1)}(x) - [\Pi_s]_{(p-1)I+1}^{(p-1)I+\ell-1} \Pi_{p-1,\ell,p-1}\left(t_p - t_{p-1}\right)\right).$$

But we see from (4.11) that $\left[\phi_{i-s}^{((p-1)I+\ell-1)}(x)\right]_{i=d_{p\ell-1}+1}^{d_{p\ell}}$ is the last s-block matrix $[\Pi_s]_{(p-1)I+1}^{(p-1)I+\ell-1} \Pi_{p-1,\ell,p-1}\Sigma_{\ell,t_{p-1}}$ of $\mathcal{S}_\ell^{(p-1,\ell-1)}$. Hence by property (4.7) of the ReLU,

$$\begin{aligned} \phi_{d_{p\ell-1}+\alpha}^{((p-1)I+\ell)}(x) &= w_s^{((p-1)I+\ell)}\sigma\Big([\Pi_s]_{(p-1)I+1}^{(p-1)I+\ell-1} \Pi_{p-1,\ell,p-1}\sigma\left(\widehat{\phi}_{d_\ell-s+\alpha}^{(I)}(x) - t_{p-1}\right) \\ &\quad - [\Pi_s]_{(p-1)I+1}^{(p-1)I+\ell-1} \Pi_{p-1,\ell,p-1}\left(t_p - t_{p-1}\right)\Big) \\ &= [\Pi_s]_{(p-1)I+1}^{(p-1)I+\ell} \Pi_{p-1,\ell,p-1}\sigma\left(\widehat{\phi}_{d_\ell-s+\alpha}^{(I)}(x) - t_p\right), \qquad \alpha = 1,\ldots,s. \end{aligned}$$

It implies $\left[\phi_i^{((p-1)I+\ell)}(x)\right]_{i=d_{p\ell-1}+1}^{d_{p\ell}} = [\Pi_s]_{(p-1)I+1}^{(p-1)I+\ell}\,\Pi_{p-1,\ell,p-1}\Sigma_{\ell,t_p}$. This proves (4.11) with $\ell-1$ replaced by ℓ. So by induction, the expression (4.11) holds true for $\ell = 2,\ldots,I+1$. Take the special index $\ell = I+1$, this expression is exactly the desired formula (4.10) for $\Phi^{(pI)}(x)$, which completes the induction procedure. □

5 Approximation Scheme for Universality

In this section we prove the universality of approximation of the DDCNN.

5.1 Mask factorization

We first need to factorize an arbitrarily fixed filter mask W supported on $\{0,1,\ldots,S_W\}$ into the convolutions of a finite sequence of filter masks $\{w^{(j)}\}_{j=1}^I$ supported on $\{0,1,\ldots,s\}$ as expressed in (3.7). For our construction, we introduce the **symbol** $\widetilde{w}$ of a filter mask $w:\mathbb{Z}\to\mathbb{R}$ supported on $\{0,1,\ldots,\tau\}$ with $\tau\in\mathbb{Z}_+$ to be a polynomial on $\mathbb{C}$ given by

$$\widetilde{w}(z) = \sum_{k=0}^{\tau} w_k z^k, \qquad z\in\mathbb{C}. \tag{5.1}$$

This concept is widely used in the literature of wavelet analysis [7]. Our key idea for proving the universality theorem here is the following factorization of the symbol $\widetilde{W}$ which is of independent interest. For $u>0$, we denote $[u]$ its integer part, and $\lceil u\rceil$ the smallest integer greater than or equal to u.

Lemma 4. *Let $s\geq 2$ and $W:\mathbb{Z}\to\mathbb{R}$ be a filter mask supported on $\{0,1,\ldots,S_W\}$ satisfying $W_0>0$ and $W_{S_W}\neq 0$ with $S_W\geq d\geq s$. Then there exist some integer $I\in\left[\frac{S_W}{s},\frac{S_W}{s-1}+2\right)$ and a sequence of filter masks $\{w^{(j)}\}_{j=1}^I$ supported on $\{0,1,\ldots,s\}$ having $w_0^{(1)}=W_0>0$ and $w_0^{(j)}=1$ for $j=2,\ldots,I$ such that the convolutional factorization (3.7) holds true.*

Proof. A useful property of the filter symbol $\widetilde{w}$ is that the symbol of the convolution $a*b$ of two finitely supported sequences a,b on $\mathbb{Z}_+$ is given by the product of the symbols of a and b as

$$\widetilde{a*b}(z) = \widetilde{a}(z)\widetilde{b}(z), \qquad \forall z\in\mathbb{C}.$$

Observe that the coefficients of the polynomial $\widetilde{W}$ of degree S_W are real. It follows that if $z_0\in\mathbb{C}$ is a root of $\widetilde{W}$ of order $\alpha\in\mathbb{N}$ meaning that its value at z_0 and the values of its derivatives up to order $\alpha-1$ vanish $\widetilde{W}(z_0)=\widetilde{W}'(z_0)=\ldots=\widetilde{W}^{(\alpha-1)}(z_0)=0$ while $\widetilde{W}^{(\alpha)}(z_0)\neq 0$, then its complex conjugate $\overline{z_0}\in\mathbb{C}$ is also a root of order α. Hence the complete factorization of the polynomial $\widetilde{W}$ has the form

$$\widetilde{W}(z) = W_{S_W}\Pi_{k=1}^A\left\{(z-z_k)(z-\overline{z_k})\right\}\Pi_{k=2A+1}^{S_W}(z-x_k), \qquad z\in\mathbb{C}, \tag{5.2}$$

where $A \in \{1, \dots, S_W\}$ is the number of (non-real) complex root pairs with multiplicity represented by $\{z_k = x_k + iy_k\}_{k=1}^{A} \subset \mathbb{C} \setminus \mathbb{R}$ with $y_k \neq 0$, $\{x_k\}_{k=2A+1}^{S_W} \subset \mathbb{R} \setminus \{0\}$ are real roots of $\widetilde{W}$ with multiplicity which are nonzero due to the assumption $\widetilde{W}(0) = W_0 > 0$. When $\widetilde{W}$ has only real roots, $A = 0$ and there is no factor of the form $(z - z_k)(z - \overline{z_k})$. When $\widetilde{W}$ has no real root, $A = \frac{S_W}{2}$ and there is no factor of the form $z - x_k$.

Note that

$$(z - z_k)(z - \overline{z_k}) = z^2 - (z_k + \overline{z_k})\, z + |z_k|^2 = z^2 - 2x_k z + \left(x_k^2 + y_k^2\right)$$

which is a quadratic polynomial with nonzero constant term due to $y_k \neq 0$. So we can normalize this quadratic polynomial to have the constant term 1 by multiplying with $1/(x_k^2 + y_k^2)$. In the same way, for $k \geq 2A + 1$, we can normalize the linear polynomial factor $z - x_k$ to have the constant term 1 by multiplying with $-1/x_k$. Thus we can factorize the polynomial $\widetilde{W}(z)$ with the constant term $W_0 > 0$ as

$$\widetilde{W}(z) = W_0 \Pi_{k=1}^{A} \left(1 - \frac{2x_k}{x_k^2 + y_k^2} z + \frac{1}{x_k^2 + y_k^2} z^2\right) \Pi_{k=2A+1}^{S_W} \left(1 - \frac{1}{x_k} z\right), \qquad z \in \mathbb{C}. \tag{5.3}$$

This leads us to construct the filter masks $\{w^{(j)}\}_{j=1}^{I}$ by their symbols as follows.

If $A \geq 1$ meaning that $\widetilde{W}$ has A non-real complex root pairs (with multiplicity), we take $I = I_1 + I_2$ with $I_1 := \lceil \frac{A}{[\frac{s}{2}]} \rceil, I_2 := \lceil \frac{S_W - 2A}{s} \rceil$ and construct the filter masks $\{w^{(j)}\}_{j=1}^{I}$ by grouping the quadratic factors in (5.3) into groups of $[s/2]$ and linear factors into groups of s as

$$\widetilde{w^{(j)}}(z) = \begin{cases} W_0 \Pi_{k=1}^{[s/2]} \left(1 - \frac{2x_k}{x_k^2+y_k^2} z + \frac{1}{x_k^2+y_k^2} z^2\right), & \text{if } j = 1, \\ \Pi_{k=(j-1)[s/2]+1}^{j[s/2]} \left(1 - \frac{2x_k}{x_k^2+y_k^2} z + \frac{1}{x_k^2+y_k^2} z^2\right), & \text{if } 2 \leq j \leq \left\lfloor \frac{A}{[\frac{s}{2}]} \right\rfloor, \\ \Pi_{k=\left\lfloor \frac{A}{[\frac{s}{2}]} \right\rfloor [\frac{s}{2}]+1}^{A} \left(1 - \frac{2x_k}{x_k^2+y_k^2} z + \frac{1}{x_k^2+y_k^2} z^2\right), & \text{if } I_1 > \frac{A}{[\frac{s}{2}]} \text{ and } j = I_1, \\ \Pi_{k=2A+1+(j-I_1-1)s}^{2A+(j-I_1)s} \left(1 - \frac{1}{x_k} z\right), & \text{if } I_1 + 1 \leq j \leq I_1 + \left\lfloor \frac{S_W - 2A}{s} \right\rfloor, \\ \Pi_{k=2A+\left\lfloor \frac{S_W-2A}{s} \right\rfloor s+1}^{S_W} \left(1 - \frac{1}{x_k} z\right), & \text{if } I_2 > \frac{S_W - 2A}{s} \text{ and } j = I. \end{cases} \tag{5.4}$$

If $A = 0$, then $\widetilde{W}$ has only S_W real roots. We can we take $I = \lceil \frac{S_W}{s} \rceil$ and construct the filter masks $\{w^{(j)}\}_{j=1}^{I}$ by grouping the linear factors into groups of s as

$$\widetilde{w^{(j)}}(z) = \begin{cases} W_0 \Pi_{k=1}^{s} \left(1 - \frac{1}{x_k} z\right), & \text{if } j = 1, \\ \Pi_{k=(j-1)s+1}^{js} \left(1 - \frac{1}{x_k} z\right), & \text{if } 2 \leq j \leq \left\lfloor \frac{S_W}{s} \right\rfloor, \\ \Pi_{k=\left\lfloor \frac{S_W}{s} \right\rfloor s+1}^{S_W} \left(1 - \frac{1}{x_k} z\right), & \text{if } \frac{S_W}{s} > \left\lfloor \frac{S_W}{s} \right\rfloor \text{ and } j = I. \end{cases} \tag{5.5}$$

With the above construction, we have

$$\widetilde{W}(z) = \Pi_{j=1}^{I} \widetilde{w^{(j)}}(z), \qquad \forall z \in \mathbb{C},$$

which yields the convolutional factorization (3.7). We also see that in our construction, $w_0^{(1)} = W_0 > 0$ and $w_0^{(j)} = 1$ for $j = 2, \ldots, I$.

Moreover, the number I of filter masks is at least $\frac{S_W}{s}$ since $I \geq \frac{A}{\frac{s}{2}} + \frac{S_W - 2A}{s} = \frac{S_W}{s}$ for the case $A \geq 1$. It can also be bounded as

$$I < \frac{A}{\lceil \frac{s}{2} \rceil} + 1 + \frac{S_W - 2A}{s} + 1 \leq \frac{2A}{s-1} + 1 + \frac{S_W - 2A}{s} + 1 \leq \frac{S_W}{s-1} + 2$$

for the case $A \geq 1$, which is also true for the case $A = 0$. This proved the desired statements. □

5.2 Polynomial factorization and features for approximation

Following our brief description in Section 2, we want to decompose an approximation $P_\Gamma \in \mathcal{P}_\Gamma(\mathbb{R}^d)$ of a function $f \in C(\Omega)$ into the form (2.1). Here we can even take the set of features $\{\xi_k\}_{k=1}^{n_\Gamma}$ to depend only on d and Γ according to the following lemma proved using some results from [19] and [20]. Denote $n_\Gamma = \begin{pmatrix} d-1+\Gamma \\ \Gamma \end{pmatrix}$ to be the dimension of $\mathcal{P}_\Gamma^h(\mathbb{R}^d)$, the space of all homogeneous polynomials on $\mathbb{R}^d$ of degree Γ.

Lemma 5. *Let $d \in \mathbb{N}$ and $\Gamma \in \mathbb{N}$. Then there exists a set $\{\xi_k\}_{k=1}^{n_\Gamma} \subset \{\xi \in \mathbb{R}^d : \|\xi\|_2 = 1\}$ of vectors with ℓ_2-norm 1 such that for any $P_\Gamma \in \mathcal{P}_\Gamma(\mathbb{R}^d)$ we can find a set of univariate polynomials $\{p_{k,\Gamma}\}_{k=1}^{n_\Gamma} \subset \mathcal{P}_\Gamma(\mathbb{R})$ which makes the identity (2.1) valid on $\mathbb{R}^d$.*

Proof. It was shown in [19] that the space $\mathcal{P}_\Gamma^h(\mathbb{R}^d)$ of homogeneous polynomials has a basis $\{(\xi_k \cdot x)^\Gamma\}_{k=1}^{n_\Gamma}$ for some vector set $\{\xi_k\}_{k=1}^{n_\Gamma} \subset \mathbb{R}^d \setminus \{0\}$. It was further proved in [20] that the vector set $\{\xi_k\}_{k=1}^{n_\Gamma} \subset \mathbb{R}^d$ can even be chosen in such way that the homogeneous polynomial set $\{(\xi_k \cdot x)^\gamma\}_{k=1}^{n_\Gamma}$ spans the space $\mathcal{P}_\gamma^h(\mathbb{R}^d)$ for every $\gamma \in \{0, 1, \ldots, \Gamma - 1\}$. Moreover, we can normalized the vectors $\{\xi_k\}_{k=1}^{n_\Gamma}$ to have ℓ_2-norm 1 since none of them is the zero vector.

The polynomial $P_\Gamma \in \mathcal{P}_\Gamma(\mathbb{R}^d)$ can be decomposed into a sum of homogeneous polynomials with various degrees as

$$P_\Gamma = \sum_{\gamma=0}^{\Gamma} P_{\gamma,\Gamma}, \quad \text{where } P_{\gamma,\Gamma} \in \mathcal{P}_\gamma^h(\mathbb{R}^d).$$

But $\{(\xi_k \cdot x)^\gamma\}_{k=1}^{n_\Gamma}$ spans $\mathcal{P}_\gamma^h(\mathbb{R}^d)$ for $\gamma \in \{0, 1, \ldots, \Gamma\}$. So there exist a set of coefficients $\{c_{k,\gamma} \in \mathbb{R}\}_{k=1}^{n_\Gamma}$ for every $\gamma \in \{0, 1, \ldots, \Gamma\}$ such that

$$P_{\gamma,\Gamma}(x) = \sum_{k=1}^{n_\Gamma} c_{k,\gamma} (\xi_k \cdot x)^\gamma, \qquad \forall\, x \in \mathbb{R}^d,\ \gamma \in \{0, 1, \ldots, \Gamma\}.$$

It follows that

$$P_\Gamma(x) = \sum_{\gamma=0}^{\Gamma} \sum_{k=1}^{n_\Gamma} c_{k,\gamma} (\xi_k \cdot x)^\gamma = \sum_{k=1}^{n_\Gamma} \left\{ \sum_{\gamma=0}^{\Gamma} c_{k,\gamma} (\xi_k \cdot x)^\gamma \right\}, \qquad \forall\ x \in \mathbb{R}^d.$$

This verifies (2.1) by setting the univariate polynomials $\{p_{k,\Gamma}\}_{k=1}^{n_\Gamma} \subset \mathcal{P}_\Gamma(\mathbb{R})$ as $p_{k,\Gamma}(u) = \sum_{\gamma=0}^{\Gamma} c_{k,\gamma} u^\gamma$. The proof of the lemma is complete. □

The univariate polynomials $\{p_{k,\Gamma}\} \subset \mathcal{P}_\Gamma(\mathbb{R})$ need to be approximated by splines $\{\sigma(u - t_j)\}_j$ with knots $\{t_j\}$. The following result is well-known in approximation theory. For completeness, we give a proof here.

Lemma 6. *Let* $\mathbf{t} := \{t_1 < t_2 \ldots < t_{N-1} < t_N\}$. *Construct a linear operator* $L_{\mathbf{t}}$ *on* $C[t_2, t_{N-1}]$ *by*

$$L_{\mathbf{t}}(f)(u) \;=\; \sum_{j=2}^{N-1} f(t_j)\delta_j(u), \qquad u \in [t_2, t_{N-1}], f \in C[t_2, t_{N-1}], \tag{5.6}$$

where the function $\delta_j \in C(\mathbb{R})$ *with* $j \in \{2, \ldots, N-1\}$ *is given by*

$$\delta_j(u) = \frac{1}{t_j - t_{j-1}}\sigma(u - t_{j-1}) - \frac{t_{j+1} - t_{j-1}}{(t_{j+1} - t_j)(t_j - t_{j-1})}\sigma(u - t_j) + \frac{1}{t_{j+1} - t_j}\sigma(u - t_{j+1}).$$

Then for any $f \in C[t_2, t_{N-1}]$, *we have*

$$\|L_{\mathbf{t}}(f) - f\|_{C[t_2, t_{N-1}]} \le 2\omega(f, \Delta_{\mathbf{t}}), \tag{5.7}$$

where $\Delta_{\mathbf{t}} := \max_{j=3,\ldots,N-1}\{|t_j - t_{j-1}|\}$ *and* $\omega(f, \mu)$ *is the modulus of continuity of* $f \in C[t_2, t_{N-1}]$ *give by*

$$\omega(f, \mu) = \sup\left\{|f(v) - f(y)| : v, y \in [t_2, t_{N-1}],\ |v - y| \le \mu\right\}, \qquad \mu > 0.$$

Proof. The function $\delta_j \in C(\mathbb{R})$ satisfies

$$\delta_j(u) = \begin{cases} \frac{u - t_{j-1}}{t_j - t_{j-1}}, & \text{if } u \in [t_{j-1}, t_j], \\ \frac{t_{j+1} - u}{t_{j+1} - t_j}, & \text{if } u \in (t_j, t_{j+1}], \\ 0, & \text{otherwise.} \end{cases}$$

It is a continuous piecewise linear function on the interval $[t_1, t_N]$ having the values $\delta_j(t_j) = 1$ and $\delta_j(t_i) = 0$ for $i \in \{1, 2, \ldots, N-1, N\} \setminus \{j\}$. This tells us that the function $L_{\mathbf{t}}(f)$ is continuous and piecewise linear on the interval $[t_2, t_{N-1}]$ with knots $\{t_2, \ldots, t_{N-1}\}$ and it interpolates f at the nodes $\{t_2, \ldots, t_{N-1}\}$. So for $j = 3, \ldots, N-1$, we have

$$L_{\mathbf{t}}(f)(u) = f(t_{j-1}) + \frac{f(t_j) - f(t_{j-1})}{t_j - t_{j-1}}(u - t_{j-1}), \qquad \forall\ u \in [t_{j-1}, t_j].$$

Hence for $u \in [t_{j-1}, t_j]$, there holds

$$\begin{aligned} |L_{\mathbf{t}}(f)(u) - f(u)| &= \left| f(t_{j-1}) - f(u) + \frac{f(t_j) - f(t_{j-1})}{t_j - t_{j-1}}(u - t_{j-1}) \right| \\ &\leq |f(t_{j-1}) - f(u)| + |f(t_j) - f(t_{j-1})|. \end{aligned}$$

Since $|t_{j-1} - u| \leq \Delta_{\mathbf{t}}$ and $|t_j - t_{j-1}| \leq \Delta_{\mathbf{t}}$, we have $|L_{\mathbf{t}}(f)(u) - f(u)| \leq 2\omega(f, \Delta_{\mathbf{t}})$ for $u \in [t_{j-1}, t_j]$ and $j = 3, \ldots, N-1$. This proves the desired error bound. □

5.3 Proof of the Main Result

We are now in a position to prove our main result on the universality of the DDCNN.

Proof of Theorem 1. Let $\epsilon > 0$. Then there exists some polynomial $P_\Gamma \in \mathcal{P}_\Gamma(\mathbb{R}^d)$ with $\Gamma \in \mathbb{N}$ such that

$$\|f - P_\Gamma\|_{C(\Omega)} \leq \frac{\epsilon}{2}.$$

By Lemma 5, there exists a vector set $\{\xi_k\}_{k=1}^{n_\Gamma} \subset \mathbb{R}^d$ such that $\|\xi_k\|_2 = 1$ for each k and the polynomial $P_\Gamma \in \mathcal{P}_\Gamma(\mathbb{R}^d)$ can be expressed as in (2.1) with $\{p_{k,\Gamma}\}_{k=1}^{n_\Gamma} \subset \mathcal{P}_\Gamma(\mathbb{R})$.

Observe that $|\xi_k \cdot x| \leq \|\xi_k\|_2 \|x\|_2 \leq \|x\|_2$ for each $k \in \{1, \ldots, n_\Gamma\}$. Hence for $x \in \Omega$, we have $|\xi_k \cdot x| \leq B_2^{(0)}$ where

$$B_2^{(0)} := \max_{x \in \Omega} \|x\|_2 < \infty.$$

So we consider the approximation of the univariate functions $\{p_{k,\Gamma}\}$ on the interval $[-B_2^{(0)}, B_2^{(0)}]$ using the spline approximation scheme in Lemma 6. For $4 \leq N \in \mathbb{N}$ we take a knot sequence $\mathbf{t} = \{t_1 < t_2 \ldots < t_{N-1} < t_N\}$ as

$$t_j = -B_2^{(0)} + (j-2)\frac{2B_2^{(0)}}{N-3}, \qquad j = 1, 2, \ldots, N-1, N$$

which implies $[t_2, t_{N-1}] = [-B_2^{(0)}, B_2^{(0)}]$ and $\Delta_{\mathbf{t}} = \frac{2B_2^{(0)}}{N-3}$. By Lemma 6,

$$\|L_{\mathbf{t}}(p_{k,\Gamma}) - p_{k,\Gamma}\|_{C[-B_2^{(0)}, B_2^{(0)}]} \leq 2\omega\left(p_{k,\Gamma}, \frac{2B_2^{(0)}}{N-3}\right).$$

Since $\lim_{\mu \to 0+} \omega(g, \mu) = 0$ for any $g \in C[-B_2^{(0)}, B_2^{(0)}]$, we know that there exists some $\mu_\epsilon > 0$ such that $\omega(p_{k,\Gamma}, \mu) \leq \frac{\epsilon}{4n_\Gamma}$ for any $0 < \mu \leq \mu_\epsilon$ and $k = 1, \ldots, n_\Gamma$. Take some $N \in \mathbb{N}$ such that $\frac{2B_2^{(0)}}{N-3} \leq \mu_\epsilon$ and $N \geq 4$. Then we have

$$\left\| \sum_{k=1}^{n_\Gamma} L_{\mathbf{t}}(p_{k,\Gamma})(\xi_k \cdot x) - P_\Gamma \right\|_{C(\Omega)} \leq \sum_{k=1}^{n_\Gamma} \|L_{\mathbf{t}}(p_{k,\Gamma})(u) - p_{k,\Gamma}(u)\|_{C[-B_2^{(0)}, B_2^{(0)}]} \leq \frac{\epsilon}{2}. \quad (5.8)$$

Now we turn to construct a sequence W (by stacking the vectors $\{\xi_k\}_{k=1}^{n_\Gamma}$ in reverse orders), which is a key step in our proof. Express each vector ξ_k in terms of its components

$$\xi_k = ((\xi_k)_\ell)_{\ell=1}^d \in \mathbb{R}^d, \qquad k = 1, \ldots, n_\Gamma.$$

Define a filter mask $W : \mathbb{Z} \to \mathbb{R}$ supported on $\{0, 1, \ldots, dn_\Gamma+1\}$ by $W_0 = W_{dn_\Gamma+1} = 1$ and

$$W_{(k-1)d+\ell} = (\xi_k)_{d+1-\ell}, \qquad \ell = 1, \ldots, d, \ k = 1, \ldots, n_\Gamma. \tag{5.9}$$

Then we apply Lemma 4 to the filter mask $W : \mathbb{Z} \to \mathbb{R}$ with $S_W = dn_\Gamma + 1$ and know that we can construct a sequence of filter masks $\{w^{(j)}\}_{j=1}^I$ supported on $\{0, 1, \ldots, s\}$ with some integer $I \in \left[\frac{dn_\Gamma+1}{s}, \frac{dn_\Gamma+1}{s-1} + 2\right)$ having $w_0^{(j)} = 1$ for $j = 1, \ldots, I$ such that the convolutional factorization (3.7) holds true. Combining this with our construction of initializing layers and the expression (3.8) for the I-th homogenized layer, we know that for $k = 1, \ldots, n_\Gamma$, the $(kd+1)$-th component of $\widehat{\phi}^{(I)}$ is

$$\widehat{\phi}_{kd+1}^{(I)}(x) = W_{kd}x_1 + \ldots + W_{kd-d+1}x_d = \sum_{\ell=1}^d W_{(k-1)d+\ell}x_{d+1-\ell} = \xi_k \cdot x. \tag{5.10}$$

By Lemma 3, there holds

$$\Phi^{(NI)}(x) = \begin{bmatrix} [\Pi_0]_{I+1}^{NI}\mathcal{L} \\ \mathcal{S}_1^{(N)} \\ \vdots \\ \mathcal{S}_I^{(N)} \end{bmatrix}, \ \mathcal{S}_j^{(N)} = \begin{bmatrix} \Pi_{N,j,1}\Sigma_{j,t_1} \\ \vdots \\ \Pi_{N,j,N}\Sigma_{j,t_N} \end{bmatrix}, \ \Sigma_{j,t} = \begin{bmatrix} \sigma\left(\widehat{\phi}_{d+(j-1)s+1}^{(I)}(x) - t\right) \\ \vdots \\ \sigma\left(\widehat{\phi}_{d+js}^{(I)}(x) - t\right) \end{bmatrix}.$$

Take $J = NI$ and $b^{(J)} \equiv 0$. Combining the above expression for $\Phi^{(J)}(x)$ and the property (4.7) of the ReLU, we find that the span of the last Js components of $\Phi^{(J)}(x)$ is

$$\operatorname{span}\left\{\Phi_i^{(J)}(x)\right\}_{i=d+1}^{d+Js} = \operatorname{span}\left\{\sigma\left(\widehat{\phi}_{d+\ell}^{(I)}(x) - t_i\right), \ \ell = 1, \ldots, Js, \ i = 1, \ldots, N\right\}.$$

Taking only those components with $\ell = (k-1)d + 1$ for $k = 1, \ldots, n_\Gamma$ and applying the identity (5.10), we have

$$\begin{aligned} & \operatorname{span}\left\{\sigma\left(\xi_k \cdot x - t_i\right), \ k = 1, \ldots, n_\Gamma, \ i = 1, \ldots, N\right\} \\ = \ & \operatorname{span}\left\{\sigma\left(\widehat{\phi}_{kd+1}^{(I)}(x) - t_i\right), \ k = 1, \ldots, n_\Gamma, \ i = 1, \ldots, N\right\} \\ \subseteq \ & \operatorname{span}\left\{\Phi_i^{(J)}(x)\right\}_{i=d+1}^{d+Js} \subseteq \mathcal{H}_J^{\mathbf{w},\mathbf{b}}. \end{aligned}$$

But the definition of the linear operator $L_{\mathbf{t}}$ tells us that

$$L_{\mathbf{t}}(p_{k,\Gamma})(\xi_k \cdot x) \in \operatorname{span}\left\{\sigma\left(\xi_k \cdot x - t_i\right), \ i = 1, \ldots, N\right\}$$

for each $k \in \{1, \ldots, n_\Gamma\}$. Hence

$$\sum_{k=1}^{n_\Gamma} L_{\mathbf{t}}(p_{k,\Gamma})(\xi_k \cdot x) \in \text{span}\left\{\sigma\left(\xi_k \cdot x - t_i\right),\ k = 1, \ldots, n_\Gamma,\ i = 1, \ldots, N\right\} \subseteq \mathcal{H}_J^{\mathbf{w},\mathbf{b}}.$$

Therefore, by taking $f^* = \sum_{k=1}^{n_\Gamma} L_{\mathbf{t}}(p_{k,\Gamma})(\xi_k \cdot x) \in \mathcal{H}_J^{\mathbf{w},\mathbf{b}}$, we obtain

$$\|f - f^*\|_{C(\Omega)} \leq \|f - P_\Gamma\|_{C(\Omega)} + \|P_\Gamma - f^*\|_{C(\Omega)} \leq \epsilon.$$

This proves the desired limit (1.12). The proof of Theorem 1 is complete. □

Acknowledgments

The work described in this paper is supported partially by the Research Grants Council of Hong Kong [Project No CityU 11306617] and by National Nature Science Foundation of China [Grant No 11461161006]. The paper was partially written when the author visited Shanghai Jiao Tong University (SJTU). The hospitality and sponsorships from SJTU and the Ministry of Education are greatly appreciated.

References

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

[2] J. Bruna and S. Mallat, Invariant scattering convolution networks, IEEE Trans. Pattern Anal. Mach. Intell. **35** (2013), 1872–1886.

[3] C. K. Chui, X. Li, H. N. Mhaskar, Neural networks for lozalized approximation, Math. Comput. **63** (1994), 607–623.

[4] C. K. Chui, X. Li, H. N. Mhaskar, Limitations of the approximation capabilities of neural networks with one hidden layer, Adv. Comput. Math. **5** (1996), 233–243.

[5] C. K. Chui, S. B. Lin, and D. X. Zhou, Construction of neural networks for realization of localized deep learning, preprint, 2018.

[6] G. Cybenko, Approximations by superpositions of sigmoidal functions, Math. Control, Signals, and Systems **2** (1989), 303–314.

[7] I. Daubechies, Ten Lectures on Wavelets, SIAM, 1992.

[8] J. Fan, T. Hu, Q. Wu and D. X. Zhou, Consistency analysis of an empirical minimum error entropy algorithm, Appl. Comput. Harmonic Anal. **41** (2016), 164–189.

[9] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.

[10] Z. C. Guo, S. B. Lin, and D. X. Zhou, Learning theory of distributed spectral algorithms, Inverse Problems **33** (2017) 074009 (29pp).

[11] Z. C. Guo, D. H. Xiang, X. Guo, and D. X. Zhou, Thresholded spectral algorithms for sparse approximations, Anal. Appl. **15** (2017), 433–455.

[12] G. E. Hinton, S. Osindero, Y. W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. **18** (2006), 1527–1554.

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

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE **86** (1998), 2278–2324.

[15] Y. LeCun, The unreasonable effectiveness of deep learning. In Seminar. Johns Hopkins University, 2014.

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

[17] H. W. Lin, M. Tegmark, and D. Rolnick, Why does deep and cheap learning work so well? J. Stat. Phys. **168** (2017), 1223-1247.

[18] S. B. Lin, X. Guo, and D. X. Zhou, Distributed learning with regularized least squares, J. Machine Learning Research **18** (92): 1–31, 2017.

[19] Y. V. Lin and A. Pinkus, Fundamentality of ridge functions, J. Approx. Theory **75** (1993), 295–311.

[20] V. Maiorov, On best approximation by ridge functions, J. Approx. Theory **99** (1999), 68–94.

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

[22] H. N. Mhaskar and T. Poggio, Deep vs. shallow networks: An approximation theory perspective, Anal. Appl. **14** (2016), 829–848.

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

[24] Y. Ying and D. X. Zhou, Unregularized online learning algorithms with general loss functions, Appl. Comput. Harmonic Anal. **42**(2017), 224–244.