UPPER BOUNDS ON THE NODE NUMBERS OF HIDDEN LAYERS IN MLPS

J. Liu[*], *F. Ni*[†], *M. Du*[†], *X. Zhang*[†], *Z. Que*[†], *S. Song*[†]

Abstract: It is one of the fundamental and challenging problems to determine the node numbers of hidden layers in neural networks. Various efforts have been made to study the relations between the approximation ability and the number of hidden nodes of some specific neural networks, such as single-hidden-layer and two-hidden-layer feedforward neural networks with specific or conditional activation functions. However, for arbitrary feedforward neural networks, there are few theoretical results on such issues. This paper gives an upper bound on the node number of each hidden layer for the most general feedforward neural networks called multilayer perceptrons (MLP), from an algebraic point of view. First, we put forward the method of expansion linear spaces to investigate the algebraic structure and properties of the outputs of MLPs. Then it is proved that given k distinct training samples, for any MLP with k nodes in each hidden layer, if a certain optimization problem has solutions, the approximation error keeps invariant with adding nodes to hidden layers. Furthermore, it is shown that for any MLP whose activation function for the output layer is bounded on $\mathbb{R}$, at most k hidden nodes in each hidden layer are needed to learn k training samples.

Key words: *feedforward neural network, hidden node, multilayer perceptron (MLP), upper bound*

Received: May 13, 2020
Revised and accepted: August 30, 2021

DOI: 10.14311/NNW.2021.31.016

1. Introduction

Neural networks can provide models for a large class of natural and artificial phenomena that are difficult to handle using classical parametric techniques. The widespread popularity of neural networks in many fields is mainly due to their ability to approximate complex nonlinear mappings directly from the input samples [1–4, 6–10, 17]. Then, a fundamental question that is often raised is how large does the network have to be to perform the approximation task. In particular,

[*] Jiang Liu; Department of Systems Science, University of Shanghai for Science and Technology, Shanghai 200093, China; School of Mathematics, University of Minnesota, Minneapolis, MN 55455, USA, E-mail: liu01306@umn.edu

[†] Feng Ni – Corresponding author; Mingjun Du; Xuyang Zhang; Zhongli Que; Shihang Song; Department of Systems Science, University of Shanghai for Science and Technology, Shanghai 200093, China, E-mail: nifeng0921@163.com

determining the optimal number of hidden nodes is one of the most challenging aspects of neural network design.

Various efforts have been made to explore the relations between the approximation ability and the number of nodes of some specific neural network, such as single-hidden-layer feedforward neural networks (SLFNs), and two-hidden-layer feedforward neural networks with specific or conditional activation functions [11–28]. For example, it was proved that N arbitrary distinct samples can be learned precisely by standard SLFNs with N hidden neurons (including biases) and the signum activation function in [12]. The bounds on the number of the hidden neurons were derived in [12] by finding particular hyperplanes that separate the input samples.

Sartori [13] presented a new method for the bounds on the size of a multilayer neural network to exactly implement an arbitrary training, which does not require the separation of the input space by particular hyperplanes, and the weights for the hidden layer can be chosen "almost" arbitrarily.

Tamura [14] pointed out that four-layered neural networks by giving any N input-target relations with a negligibly small error using $\frac{N}{2}+3$ hidden units and sigmoid activation. Huang [15] later extended the work of Tamura and Tateshi to prove that the upper bound on the number of hidden nodes N_h for TLFNs with sigmoid activation function is given by $N_\mathrm{h} \leq 2\sqrt{(N_0+2)N_\mathrm{s}}$, where N_0 is the number of outputs, and N_s is the number of samples.

Huang [16] showed that an SLFN with at most N hidden nodes and with any arbitrary bounded nonlinear activation function which has a limit at one infinity can exactly learn N distinct observations.

Later, Huang [17] proved that if the number of hidden nodes is equal to the number of distinct training samples, SLFNs with random input weight vectors and hidden biases can approximate the training samples with zero error. Furthermore, it was proved by [18–20] that for SLFNs, the approximation error is monotonically decreasing with gradually adding nodes in hidden layer.

However, for arbitrary neural networks, there are few theoretical results on the relations of approximation ability and hidden node number in the literature.

In this paper, we consider the most popular and general feedforward neural networks called multilayer perceptrons (MLP). Since in real applications, neural networks are trained using finite input samples, we focus on the approximation capabilities of multilayer feedforward neural networks approximation in a finite set of training samples. More precisely, we consider the question as follows: Given k training samples, for an s-hidden-layer MLP, how many nodes in each hidden layer are needed to learn the samples? Here s is an arbitrary positive integer.

To answer the question, we first answer the fundamental questions from the mathematical point of view: Given k training samples, what is the algebraic structure and what are the algebraic properties of the outputs of MLPs? In this paper, the output vectors of an MLP are described by the pre-output combination vector, and it is proved that the set of all possible pre-output combination vectors of an MLP by adjusting the weights and bias is a union of expansion linear spaces. Besides, the set is shown to have the property of keeping growing with the increase of node numbers until the node number in each hidden layer is k.

Finally, based on the algebraic properties, we find that given k distinct training samples, for any s-hidden-layer MLP with k nodes in each hidden layer, if a certain optimization problem has solutions, then adding nodes to hidden layers will not change the approximation error. Furthermore, it is shown that for an s-hidden-layer MLP whose activation function for the output layer is bounded on $\mathbb{R}$, at most k hidden nodes in each hidden layer are needed to learn k training samples.

2. Preliminaries and notations

A neural network consists of a number of interconnected neurons. Each neuron is a simple processing element that responds to the weighted inputs it received from other neurons. A multilayer perceptron is the most popular and general feedforward neural network. It is composed of three typical classes of layers: An input layer, that serves to pass the input vector to the network, hidden layers of computation neurons, and an output layer composed of at least one computation neuron to produce the output vector.

In this paper, it is a convention that for any s-hidden-layer MLP (or called $s+2$ layered MLP, with $s \geq 1$), layer 0 denotes the input layer, and layer $s+1$ denotes the output layer, $f_q(\cdot)$ is the activation function which applies to all neurons in layer q ($0 \leq q \leq s+1$). In particular, f_0 is an identity mapping.

Notation 1. *$\mathcal{M}_{s,[f]}^{\{e_0,e_{s+1}\}}$ denotes the set of $(s+2)$-layered MLPs in which the activation function for layer q is $f_q(\cdot)$, and the node numbers of input layer and output layer are e_0 and e_{s+1} respectively, where $0 \leq q \leq s+1$.*

$\mathcal{M}_{s,[f]}^{[e]}$ denotes the $(s+2)$-layered MLP in which the activation function for layer q is $f_q(\cdot)$, and the node number of layer q is e_q, where $0 \leq q \leq s+1$.

For layer $q \geq 1$, $\mathbf{W}^{(q)} = [w_{ij}^{(q)}]_{e_q \times e_{q-1}}$ denotes the $e_q \times e_{q-1}$ weight matrix, and $\mathbf{b}_q = (b_{q1}, \ldots, b_{qe_q})^{\mathrm{T}}$ denotes the bias vector.

Definition 1. *Suppose Γ is an $(s+2)$-layered MLP in $\mathcal{M}_{s,[f]}^{\{e_0,e_{s+1}\}}$, and $\{(\mathbf{u}_i, \mathbf{t}_i)\}_{i=1}^{k} \subseteq \mathbb{R}^{e_0} \times \mathbb{R}^{e_{s+1}}$ is a set of k training samples, where $\mathbf{u}_i \in \mathbb{R}^{e_0}$ is an input vector and $\mathbf{t}_i \in \mathbb{R}^{e_{s+1}}$ is the desired output vector. Suppose $\mathbf{o}_i^{(j)}$ is the output vector of layer $j-1$ with respect to input $\mathbf{u}_i$ ($j \geq 1$, and in particular, $\mathbf{o}_i^{(1)}$ is $\mathbf{u}_i$). Then we denote $\mathbf{W}^{(j)}\mathbf{o}_i^{(j)} + \mathbf{b}_j$ by $\mathbf{d}_{j,i}$, and call it the pre-output vector of layer j. The vector $(\mathbf{d}_{j,1}{}^{\mathrm{T}}, \mathbf{d}_{j,2}{}^{\mathrm{T}}, \ldots, \mathbf{d}_{j,k}{}^{\mathrm{T}})^{\mathrm{T}}$ is called the pre-output combination vector of layer j.*

3. Expansion linear space

In this section, we define expansion linear spaces, and provide the necessary and sufficient condition for a vector belonging to an expansion linear space, as a preparation for exploring the algebraic structure of outputs of MLPs in the next section.

In what follows, we always assume that $\mathbf{a}_1, \ldots, \mathbf{a}_m$ are column vectors in $\mathbb{R}^k$, and $k, m \in \mathbb{N}^*$.

Definition 2 (Expansion Vector List). *Suppose for $1 \leq i \leq m$, $\mathbf{a}_i = (a_{1i}, \ldots, a_{ki})^{\mathrm{T}}$. First, for each $\mathbf{a}_i$, define its p-degree expansion vector list $Bon^p(\mathbf{a}_i)$ by listing the column vectors of the following $kp \times p$ matrix from left to right. In the* *following matrix, each submatrix formed by the elements in the $sp+1$-th, $sp+2$-th, $\dots$, $sp+p$-th rows is a $p \times p$ diagonal matrix for $s = 0, 1, \dots, k-1$.*

$$\begin{pmatrix}
a_{1i} & & & \\
& a_{1i} & & \\
& & \ddots & \\
& & & a_{1i} \\
a_{2i} & & & \\
& a_{2i} & & \\
& & \ddots & \\
& & & a_{2i} \\
\vdots & \vdots & & \vdots \\
a_{ki} & & & \\
& a_{ki} & & \\
& & \ddots & \\
& & & a_{ki}
\end{pmatrix}$$

Secondly, define the p-degree expansion vector list of $\{\mathbf{a}_1, \dots, \mathbf{a}_m\}$ by listing the elements of $Bon^p(\mathbf{a}_1)$, $\dots$, $Bon^p(\mathbf{a}_m)$ in sequence, denoted by $Bon^p(\mathbf{a}_1, \dots, \mathbf{a}_m)$.

Remark 1. *For $1 \le i, j \le m$, let $f_i = a_{1i}x^{kp-p} + a_{2i}x^{kp-2p} \cdots + a_{ki}x^0$ and $f_j = a_{1j}x^{kp-p} + a_{2j}x^{kp-2p} \cdots + a_{kj}x^0$ be two polynomials in $\mathbb{R}[x]$. Then the p-degree expansion vector list of $\{\mathbf{a}_i, \mathbf{a}_j\}$ is exactly $Syl(f_i, f_j, x)^{\mathrm{T}}$, where $Syl(f_i, f_j, x)$ is the Sylvester matrix of f_i, f_j.*

Example 1. The 2-degree expansion vector list of $\{(3,1)^{\mathrm{T}}, (2,4)^{\mathrm{T}}\}$ is

$$\{(3,0,1,0)^{\mathrm{T}}, (0,3,0,1)^{\mathrm{T}}, (2,0,4,0)^{\mathrm{T}}, (0,2,0,4)^{\mathrm{T}}\}.$$

Proposition 1. *Let $\mathcal{V}$ be the linear space spanned by $\mathbf{a}_1, \dots, \mathbf{a}_m$ over $\mathbb{R}$. Suppose that $\mathcal{V}$ can be also spanned by $\mathbf{b}_1, \dots, \mathbf{b}_{m'}$. Then $Bon^p(\mathbf{a}_1, \dots, \mathbf{a}_m)$ spans the same linear space as $Bon^p(\mathbf{b}_1, \dots, \mathbf{b}_{m'})$ does, and the dimension of the space is p times the dimension of $\mathcal{V}$.*

Proof. First, directly from the definition of expansion vector list, we get the following statement: Let $\mathbf{e}_1, \dots, \mathbf{e}_q$ be vectors in $\mathcal{V}$. If they are linearly independent, all the vectors in the expansion vector list $\mathrm{Bon}^p(\mathbf{e}_1, \dots, \mathbf{e}_q)$ are linearly independent. If they are linearly dependent, without loss of generality, assuming that $\mathbf{e}_1 = \sum\limits_{i=2}^{q} d_i \mathbf{e}_i$, then $\mathbf{g}_{1j} = \sum\limits_{i=2}^{q} d_i \mathbf{g}_{ij}$, where $d_i \in \mathbb{R}$, and $\mathbf{g}_{ij}$ denotes the j-th column vector of $\mathrm{Bon}^p(\mathbf{e}_i)$, $1 \le i \le q, 1 \le j \le p$. In other words, each vector in $\mathrm{Bon}^p(\mathbf{e}_1)$ is a linear combination of vectors in $\mathrm{Bon}^p(\mathbf{e}_2, \dots, \mathbf{e}_q)$.

Since $\mathbf{a}_i$ is a linear combination of $\mathbf{b}_1, \dots, \mathbf{b}_{m'}$, each vector in $\mathrm{Bon}^p(\mathbf{a}_i)$ is a linear combination of vectors in $\mathrm{Bon}^p(\mathbf{b}_1, \dots, \mathbf{b}_{m'})$. Similarly, each vector in $\mathrm{Bon}^p(\mathbf{b}_i)$ is a linear combination of vectors in $\mathrm{Bon}^p(\mathbf{a}_1, \dots, \mathbf{a}_m)$. Therefore $\mathrm{Bon}^p(\mathbf{a}_1, \dots, \mathbf{a}_m)$ spans the same linear space as $\mathrm{Bon}^p(\mathbf{b}_1, \dots, \mathbf{b}_{m'})$ does.

Besides, from above, we conclude that for any vectors $\mathbf{e}_1, \ldots, \mathbf{e}_q$, they are linearly independent if and only if the vectors in $\mathrm{Bon}^p(\mathbf{e}_1, \ldots, \mathbf{e}_q)$ are linearly independent. This implies that the dimension of the space spanned by $\mathrm{Bon}^p(\mathbf{a}_1, \ldots, \mathbf{a}_m)$ is p times the dimension of $\mathcal{V}$. □

By Proposition 1, we can define the unique p-degree expansion linear space for any finite-dimensional linear space, as follows.

Definition 3 (Expansion Linear Space). *Suppose $\mathcal{V}$ is a finite-dimensional linear space, which can be spanned by $\mathbf{a}_1, \ldots, \mathbf{a}_m$. Then Span(Bon$^p(\mathbf{a}_1, \ldots, \mathbf{a}_m)$) is called the p-degree expansion linear space of $\mathcal{V}$, and denoted by Bon$^p(\mathcal{V})$.*

Definition 4 (Rotation Vector List). *The rotation vector list of $\{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$, denoted by Ro$(\mathbf{a}_1, \ldots, \mathbf{a}_m)$, is obtained by listing the column vectors of the matrix $(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m)^{\mathrm{T}}$ from left to right. Each column vector in Ro$(\mathbf{a}_1, \ldots, \mathbf{a}_m)$ is called a ro-column-vector of the vector $({\mathbf{a}_1}^{\mathrm{T}}, {\mathbf{a}_2}^{\mathrm{T}}, \ldots, {\mathbf{a}_m}^{\mathrm{T}})^{\mathrm{T}}$.*

From the definition of rotation vector list, the following statement is straightforward.

Lemma 1. *Suppose the column vectors in Ro$(\mathbf{a}_1, \ldots, \mathbf{a}_m)$ are $\mathbf{v}_1, \ldots, \mathbf{v}_k$. Then $\mathbf{a}_1, \ldots, \mathbf{a}_m$ must be in Ro$(\mathbf{v}_1, \ldots, \mathbf{v}_k)$.*

Lemma 2. *Let $\mathbf{d}_1, \ldots, \mathbf{d}_k$ be s-dimensional column vectors. Then the $k \times s$-dimensional vector $\mathbf{x} = ({\mathbf{d}_1}^{\mathrm{T}}, {\mathbf{d}_2}^{\mathrm{T}}, \ldots, {\mathbf{d}_k}^{\mathrm{T}})^{\mathrm{T}}$ belongs to Bon^s(Span$(\mathbf{a}_1, \ldots, \mathbf{a}_m)$) if and only if each column vector in Ro$(\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_k)$ belongs to Span$(\mathbf{a}_1, \ldots, \mathbf{a}_m)$.*

Proof. Suppose for $1 \leq i \leq m$, $\mathbf{a}_i = (a_{1i}, \ldots, a_{ki})^{\mathrm{T}}$.

Assume that $\mathbf{x}$ is in $\mathrm{Bon}^s(\mathrm{Span}(\mathbf{a}_1, \ldots, \mathbf{a}_m))$. Then $\mathbf{x}$ is a linear combination of the vectors in $\mathrm{Bon}^s(\mathbf{a}_1), \ldots, \mathrm{Bon}^s(\mathbf{a}_m)$, and we use p_{ji} to denote the coefficient in front of the i-th vector in $\mathrm{Bon}^s(\mathbf{a}_j)$. Hence, $\mathbf{d}_i = \left(\sum_{j=1}^m p_{j1} a_{ji}, \sum_{j=1}^m p_{j2} a_{ji}, \ldots, \sum_{j=1}^m p_{js} a_{ji}\right)^{\mathrm{T}}$. Furthermore, for $1 \leq i \leq s$, the i-th column vector in $\mathrm{Ro}(\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_k)$ is

$$\left(\sum_{j=1}^m p_{ji} a_{j1}, \sum_{j=1}^m p_{ji} a_{j2}, \ldots, \sum_{j=1}^m p_{ji} a_{jk}\right)^{\mathrm{T}},$$

which obviously belongs to $\mathrm{Span}(\mathbf{a}_1, \ldots, \mathbf{a}_m)$.

Conversely, assume that each column vector in $\mathrm{Ro}(\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_k)$ is in $\mathrm{Span}(\mathbf{a}_1, \ldots, \mathbf{a}_m)$, and the i-th column vector in $\mathrm{Ro}(\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_k)$ is $\sum\limits_{j=1}^{m} p_{ji} \mathbf{a}_j$. Then by a direct computation, $\mathbf{x}$ is is a linear combination of the vectors in $\mathrm{Bon}^s(\mathbf{a}_1, \ldots, \mathbf{a}_m)$, and the coefficient in front of the j-th vector in $\mathrm{Bon}^s(\mathbf{a}_i)$ is p_{ij}. □

4. Upper bounds on the node numbers

By Definition 1, the pre-output combination vector can be used to describe the outputs of an MLP with respect to given inputs. Hence, in this section, we investigate the algebraic structure and properties for the set of all possible pre-output combination vectors of an MLP by adjusting the weights and bias. Given k distinct training samples, the set is shown to be a union of some expansion linear spaces, and keep growing with the increase of node numbers until the node number in each hidden layer is k. Then, based on the algebraic properties, we prove that at most k nodes in each hidden layer are needed to learn k training samples when the activation function for the output layer is bounded on $\mathbb{R}$.

In what follows, we will use these notations frequently.

Notation 2. *(1) We denote the k-dimensional vector $(1, 1, \ldots, 1)^{\mathrm{T}}$ by $(\mathbf{1})_k$. Suppose $\mathcal{G} \subseteq \mathbb{R}^k$. $\mathcal{G} \bigcup (\mathbf{1})_k$ is denoted by $\overline{\mathcal{G}}$.*

(2) Let $\mathcal{G}$ be a subset of $\mathbb{R}^k$. For $i \geqslant 1$, we use δ_i to denote an operation of choosing i vectors from $\mathcal{G}$, all such operations consist a set denoted by Δ_i, where i is less than the element number of $\mathcal{G}$.

(3) Let $2^{\mathbb{R}^k}$ be the power set of $\mathbb{R}^k$, i.e., the set of all subsets of $\mathbb{R}^k$. For any mapping $f : 2^{\mathbb{R}^k} \longrightarrow 2^{\mathbb{R}^k}$ and $\mathcal{G} \subseteq \mathbb{R}^k$, Span$(\overline{\delta_i f(\mathcal{G})})$ is also a subset of $\mathbb{R}^k$, hence, Span$(\overline{\delta_i f(\cdot)})$ is a mapping of $2^{\mathbb{R}^k}$ into $2^{\mathbb{R}^k}$. Denote Span$(\overline{\delta_i f(\cdot)})$ by $\Psi_f^{\delta_i}(\cdot)$.

Similarly, denote the mapping Span$(\overline{f(\cdot)})$ by $\Theta_f(\cdot)$.

Example 2. Suppose $\mathcal{G}$ is $\{(2,3)^{\mathrm{T}}, (4,1)^{\mathrm{T}}, (5,9)^{\mathrm{T}}\} \subseteq \mathbb{R}^2$. Then $\delta_1\mathcal{G}$ can be $(2,3)^{\mathrm{T}}$ or $(4,1)^{\mathrm{T}}$ or $(5,9)^{\mathrm{T}}$, choosing one vector from $\mathcal{G}$.

The following proposition shows that a pre-output combination vector set is a union of linear subspaces. The proof is given in Appendix.

Proposition 2 (Algebraic Structure). *Given k input vectors $\{\mathbf{u}_i\}_{i=1}^k \subseteq \mathbb{R}^{e_0}$, all the j-layer pre-output combination vectors of $\mathcal{M}_{s,[f]}^{[e]}$ consist a set denoted by $OP_j^{[e]}$, $1 \le j \le s+1$. Then $OP_j^{[e]}$ is a union of linear subspaces, as follows:*

$$\bigcup_{\delta_i \in \Delta_i} Bon^{e_j}\left(\Psi_{f_{j-1}}^{\delta_{e_{j-1}}} \circ \cdots \circ \Psi_{f_0}^{\delta_{e_0}}\left(Ro\left(\mathbf{u}_1, \ldots, \mathbf{u}_k\right)\right)\right),$$

denoted by $\bigcup_{\delta_i \in \Delta_i} Bon^{e_j}\left(\prod_{i=0}^{j-1} \Psi_{f_i}^{\delta_{e_i}}\left(Ro\left(\mathbf{u}_1, \ldots, \mathbf{u}_k\right)\right)\right)$, where $\circ$ denotes the composition of mappings.

We directly have the following statement from Proposition 2.

Corollary 1. *Given k input vectors $\{\mathbf{u}_i\}_{i=1}^k \subseteq \mathbb{R}^{e_0}$, all the j-layer pre-output combination vectors of MLPs in $\mathcal{M}_{s,[f]}^{\{e_0,e_{s+1}\}}$ consist a set denoted by $OP_j^{\{e_0,e_j\}}$, $1 \le j \le s+1$. Then*

$$OP_j^{\{e_0,e_j\}} = Bon^{e_j}\left(\Theta_{f_{j-1}} \circ \cdots \circ \Theta_{f_0}(Ro(\mathbf{u}_1, \ldots, \mathbf{u}_k))\right),$$

which is a linear space.

The pre-output combination vectors have the following properties.

Proposition 3 (Algebraic Property). *Given k input vectors $\{\mathbf{u}_i\}_{i=1}^{k} \subseteq \mathbb{R}^{e_0}$, and $j \geq 1$.*

(1) If $e_i \leq \tilde{e}_i$ for $1 \leq i \leq j-1$ and $e_j = \tilde{e}_j$, then $OP_j^{[e]} \subseteq OP_j^{[\tilde{e}]}$.

(2) If $\tilde{e}_1, \ldots, \tilde{e}_{j-1} \geq e_1, \ldots, e_{j-1} \geq k$ and $e_j = \tilde{e}_j$, then $OP_j^{[e]}$ is a linear space, and

$$OP_j^{[e]} = OP_j^{[\tilde{e}]} = OP_j^{\{e_0, e_j\}},$$

Proof. (1) Straightforward from Proposition 2.

(2) It suffices to prove that $\mathrm{OP}_j^{\{e_0,e_j\}} \subseteq \mathrm{OP}_j^{[e]}$.

Since the vectors contained in $\overline{f_1(\mathrm{Span}(\overline{\mathrm{Ro}(\mathbf{u}_1, \ldots, \mathbf{u}_k)}))}$ are k-dimensional, at most k vectors are linearly independent. If $e_1 \geq k$, there exists $\delta_{e_1} \in \Delta_{e_1}$, such that in $\overline{f_1(\mathrm{Span}(\overline{\mathrm{Ro}(\mathbf{u}_1, \ldots, \mathbf{u}_k)}))}$, a set of vectors which are maximally linearly independent is contained in $\delta_{e_1} f_1(\mathrm{Span}(\overline{\mathrm{Ro}(\mathbf{u}_1, \ldots, \mathbf{u}_k)}))$. Therefore,

$$\begin{aligned}&\mathrm{Span}(\overline{\delta_{e_1} f_1(\mathrm{Span}(\overline{\mathrm{Ro}(\mathbf{u}_1, \ldots, \mathbf{u}_k)}))}) \\ = {}&\mathrm{Span}(\overline{f_1(\mathrm{Span}(\overline{\mathrm{Ro}(\mathbf{u}_1, \ldots, \mathbf{u}_k)}))}),\end{aligned}$$

i.e.,

$$\Theta_{f_1} \circ \Theta_{f_0}(\mathrm{Ro}(\mathbf{u}_1, \ldots, \mathbf{u}_k)) = \Psi_{f_1}^{e_1} \circ \Psi_{f_0}^{e_0}(\mathrm{Ro}(\mathbf{u}_1, \ldots, \mathbf{u}_k)).$$

Analogously, there exist $\delta_{e_2} \in \Delta_{e_2}, \delta_{e_3} \in \Delta_{e_3}, \ldots, \delta_{e_{j-1}} \in \Delta_{e_{j-1}}$, such that

$$\Theta_{f_{j-1}} \circ \cdots \circ \Theta_{f_0}(\mathrm{Ro}(\mathbf{u}_1, \ldots, \mathbf{u}_k)) = \Psi_{f_{j-1}}^{e_{j-1}} \circ \cdots \circ \Psi_{f_0}^{e_0}(\mathrm{Ro}(\mathbf{u}_1, \ldots, \mathbf{u}_k)),$$

implying

$$\mathrm{Bon}^{e_j}(\Theta_{f_{j-1}} \circ \cdots \circ \Theta_{f_0}(\mathrm{Ro}(\mathbf{u}_1, \ldots, \mathbf{u}_k))) = \mathrm{Bon}^{e_j}(\Psi_{f_{j-1}}^{e_{j-1}} \circ \cdots \circ \Psi_{f_0}^{e_0}(\mathrm{Ro}(\mathbf{u}_1, \ldots, \mathbf{u}_k))).$$

Hence, $\mathrm{OP}_j^{\{e_0,e_j\}} \subseteq \mathrm{OP}_j^{[e]}$ by Corollary 1 and Proposition 2. □

The following lemma can be easily verified by the definition of Euclidean norm of a vector.

Lemma 3. *Suppose $\mathbf{c}_i$, $\mathbf{t}_i$ $(i = 1, \ldots, k)$ are column vectors of the same dimension, then*

$$\sum_{i=1}^{k} \|\mathbf{t}_i - \mathbf{c}_i\|_2^2 = \|(\mathbf{t}_1^{\mathrm{T}}, \mathbf{t}_2^{\mathrm{T}}, \ldots, \mathbf{t}_k^{\mathrm{T}})^{\mathrm{T}} - (\mathbf{c}_1^{\mathrm{T}}, \mathbf{c}_2^{\mathrm{T}}, \ldots, \mathbf{c}_k^{\mathrm{T}})^{\mathrm{T}}\|_2^2.$$

Remark 2. *Suppose $\{(\mathbf{u}_i, \mathbf{t}_i)\}_{i=1}^{k} \subseteq \mathbb{R}^{e_0} \times \mathbb{R}^{e_{s+1}}$ is a set of training samples. Lemma 3 implies that if $\min\limits_{\mathbf{d} \in OP_{s+1}^{[e]}} \|\mathbf{t} - f_{s+1}(\mathbf{d})\|_2^2$ exists, the approximation error of $\mathcal{M}_{s,[f]}^{[e]}$ equals to $\min\limits_{\mathbf{d} \in OP_{s+1}^{[e]}} \|\mathbf{t} - f_{s+1}(\mathbf{d})\|_2^2$, where $\mathbf{t}$ denotes $(\mathbf{t}_1^{\mathrm{T}}, \ldots, \mathbf{t}_k^{\mathrm{T}})^{\mathrm{T}}$.*

The following relation between node numbers and approximation errors is straightforward from Proposition 3.

Theorem 1. *Suppose* $\{(\mathbf{u}_i,\mathbf{t}_i)\}_{i=1}^k \subseteq \mathbb{R}^{e_0} \times \mathbb{R}^{e_{s+1}}$ *is a set of training samples. Then*

$$\min_{\mathbf{d}\in OP_{s+1}^{[e^{(0)}]}} \|\mathbf{t} - f_{s+1}(\mathbf{d})\|_2 \geq \min_{\mathbf{d}\in OP_{s+1}^{[e^{(1)}]}} \|\mathbf{t} - f_{s+1}(\mathbf{d})\|_2,$$

and

$$\min_{\mathbf{d}\in OP_{s+1}^{[e^{(2)}]}} \|\mathbf{t} - f_{s+1}(\mathbf{d})\|_2 = \min_{\mathbf{d}\in OP_{s+1}^{[e^{(3)}]}} \|\mathbf{t} - f_{s+1}(\mathbf{d})\|_2,$$

where for $1 \leq i \leq s$, $e_i^{(0)} \leq e_i^{(1)}$, $e_i^{(2)} \geq k$, $e_i^{(3)} = k$, *and* $\mathbf{t}$ *denotes* $(\mathbf{t}_1^{\mathrm{T}}, \ldots, \mathbf{t}_k^{\mathrm{T}})^{\mathrm{T}}$.

Remark 3. *Theorem 1 and Remark 2 show that given* k *distinct training samples, for* $\mathcal{M}_{s,[f]}^{[e]}$ *with* $e_1 = e_2 = \cdots = e_s = k$, *if* $\min_{\mathbf{d}\in OP_{s+1}^{[e]}} \|\mathbf{t} - f_{s+1}(\mathbf{d})\|_2$ *exists, adding nodes to hidden layers will not change the approximation error.*

From Remark 3, a natural question arises: When does $\min_{\mathbf{d}\in OP_{s+1}^{[e]}} \|\mathbf{t} - f_{s+1}(\mathbf{d})\|_2$ exist? In the rest of the paper, we will show that if the activation function f_{s+1} for the output layer is bounded on $\mathbb{R}$, then $\min_{\mathbf{d}\in OP_{s+1}^{[e]}} \|\mathbf{t} - f_{s+1}(\mathbf{d})\|_2$ exists.

First, the topological property of the pre-output combination vector set $f_{s+1}(\mathrm{OP}_s^{[e]})$ is presented as follows.

Lemma 4. *Suppose* $\{(\mathbf{u}_i,\mathbf{t}_i)\}_{i=1}^k \subseteq \mathbb{R}^{e_0} \times \mathbb{R}^{e_{s+1}}$ *is a set of training samples. If the activation function* f_{s+1} *for the output layer is bounded on* $\mathbb{R}$, *then* $f_{s+1}(OP_s^{[e]})$ *is a sequentially compact set, i.e., there exists a convergence subsequence in* $f_{s+1}(OP_s^{[e]})$.

Proof. Since f_{s+1} is bounded, $f_{s+1}(\mathrm{OP}_s^{[e]})$ is a bounded subset in $\mathbb{R}^{ke_{s+1}}$. Then by Bolzano-Weierstrass Theorem (see e.g., [29]), every bounded infinite set in $\mathbb{R}^{ke_{s+1}}$ has a convergence subsequence, which completes the proof. □

Finally, it will be shown that if f_{s+1} is bounded, then $\min_{\mathbf{d}\in OP_{s+1}^{[e]}} \|\mathbf{t} - f_{s+1}(\mathbf{d})\|_2$ exists.

Lemma 5. *Given* $\{(\mathbf{u}_i,\mathbf{t}_i)\}_{i=1}^k \subseteq \mathbb{R}^{e_0} \times \mathbb{R}^{e_{s+1}}$, *define the function* $h : f_{s+1}(OP_s^{[e]}) \to \mathbb{R}$,

$$h(\mathbf{x}) = \|\mathbf{x} - \mathbf{t}\|,$$

where $\mathbf{t} = ({\mathbf{t}_1}^{\mathrm{T}}, \ldots, {\mathbf{t}_k}^{\mathrm{T}})^{\mathrm{T}}$. *If* f_{s+1} *is bounded, then* h *is a continuous bounded function, and there exists* $\mathbf{x}_0 \in f_{s+1}(OP_s^{[e]})$, *such that* $h(\mathbf{x}_0) = \inf_{\mathbf{x}\in f_{s+1}(OP_s^{[e]})} h(\mathbf{x})$.

Proof. $h(\mathbf{x})$ is bounded, since $|h(\mathbf{x})| \leq \|\mathbf{x}\| + \|\mathbf{t}\|$ and $f_{s+1}(\mathrm{OP}_s^{[e]})$ is a bounded set. Besides, $|h(\mathbf{x}_1) - h(\mathbf{x}_2)| \leq \|(\mathbf{x}_1 - \mathbf{t}) - (\mathbf{x}_2 - \mathbf{t})\| = \|\mathbf{x}_1 - \mathbf{x}_2\|$. We claim that $h(\mathbf{x})$ is continuous for the following reason: for any $\mathbf{x} \in \mathrm{OP}_s^{[e]}$ and $\varepsilon > 0$, let $\delta = \varepsilon$, when $\mathbf{x}' \in \mathrm{OP}_s^{[e]}$ and $\|\mathbf{x} - \mathbf{x}'\| < \delta$, we have $|h(\mathbf{x}) - h(\mathbf{x}')| \leq \|\mathbf{x} - \mathbf{x}'\| < \varepsilon$. Note that $f_{s+1}(\mathrm{OP}_s^{[e]})$ is sequentially compact by Lemma 4. Therefore the statement holds according to Theorem 1.2 of Chapter 2 in [30], for which the proof is following.

Let $c = \inf_{\mathbf{x}\in f_{s+1}(\mathrm{OP}_s^{[e]})} h(\mathbf{x})$, then there is a sequence $\{\mathbf{x}_n\} \subset f_{s+1}(\mathrm{OP}_s^{[e]})$, such that $\lim_{n\to+\infty} h(\mathbf{x}_n) \to c$. Since $f_{s+1}(\mathrm{OP}_s^{[e]})$ is sequentially compact, there exists a convergence subsequence $\{\mathbf{x}_{n_k}\}$. Let $\mathbf{x}_0$ be the limit of $\{\mathbf{x}_{n_k}\}$. Since h is continuous, $c = \lim_{k\to+\infty} h(\mathbf{x}_{n_k}) = h(\mathbf{x}_0)$, implying $h(\mathbf{x}_0) = \inf_{\mathbf{x}\in f_{s+1}(\mathrm{OP}_s^{[e]})} h(\mathbf{x})$. □

Theorem 2. *Lemma 5 and Theorem 1 imply that given k distinct training samples, for any MLP, if the activation function for the output layer is bounded on $\mathbb{R}$, then the approximation error keeps decreasing with the increase of node numbers until the node number in each hidden layer is k, i.e., at most k nodes in each hidden layer are needed to learn k training samples.*

Remark 4. *Parts of the conclusions presented in [12, 16] are the special cases of Theorem 2, i.e,. in the single-hidden-layer feedforward neural networks (SLFN) with the signum activation function or bounded nonlinear activation function which has a limit at one infinity, at most k hidden nodes are needed to learn k training samples.*

Conclusion

An upper bound on the node number of each hidden layer for MLPs has been derived by algebraic methods, such as the method of expansion linear spaces. Meanwhile, the basic algebraic structure and properties of the outputs of MLPs are presented. First, we use the pre-output combination vector to describe the outputs of an MLP. Then it is proved that given k input vectors, the set of all possible pre-output combination vectors of an MLP by adjusting the weights and bias is a union of expansion linear spaces. Besides, the set is found to keep growing with the increase of node numbers until the node number in each hidden layer is k. Finally, we reach the conclusions that given k training samples, for any MLP with k nodes in each hidden layer, if the activation function for the output layer is bounded on $\mathbb{R}$, then $\min_{\mathbf{d}\in\mathrm{OP}_{s+1}^{[e]}} \|\mathbf{t} - f_{s+1}(\mathbf{d})\|_2$ exists, which is exactly the approximation error and keeps invariant with the increase of node numbers.

This paper provides a general theoretical criterion for determining the hidden node numbers of MLPs from the perspective of minimizing the mean square distance between outputs and desired outputs, without considering certain issues such as convergence speed improvement and over-fitting problem arising from some certain practical applications.

The algebraic approach (especially Proposition 2) can be further used to compare approximation errors of MLPs with different numbers of layers and may help to design layer numbers for MLPs.

Acknowledgement

The authors wish to thank the reviewers and the editors for their constructive suggestions and comments. This research was supported by National Natural Science Foundation of China under Grant No. 11701370.

References

[1] HORNIK K. Approximation capabilities of multilayer feedforward networks. *Neural Networks.* 1991, 4, pp. 251–257, doi: 10.1016/0893-6080(91)90009-t.

[2] LESHNO M., LIN V.Y., PINKUS A., SCHOCKEN S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks.* 1993, 6, pp. 861–867, doi: 10.1016/s0893-6080(05)80131-5.

[3] CHEN T., CHEN H. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Trans. Neural Networks.* 1995, 6, pp. 911–917, doi: 10.1109/72.392253.

[4] STINCHCOMBE M.B. Neural networks approximation of continuous functional and continuous functions on compactifications. *Neural Netw.* 1999, 12(3), pp. 467–477, doi: 10.1016/s0893-6080(98)00108-7.

[5] HUANG G.B., CHEN L., SIEW C.K. Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Trans. Neural Netw.* 2006, 17(4), pp. 879–892, doi: 10.1109/tnn.2006.875977.

[6] ANASTASSIOU G.A. Multivariate sigmoidal neural network approximation. *Neural Netw.* 2011, 24(4), pp. 378–386, doi: 10.1016/j.neunet.2011.01.003.

[7] ZHANG R., LAN Y., HUANG G.B., XU Z.B. Universal approximation of extreme learning machine with adaptive growth of hidden nodes. *IEEE Trans. Neural Netw. Learn. Syst.* 2012, 23(2), pp. 365–371, doi: 10.1109/tnnls.2011.2178124.

[8] ANDRAS P. Function approximation using combined unsupervised and supervised learning. *IEEE Trans. Neural Netw. Learn. Syst.* 2014, 25(3), pp. 495–505, doi: 10.1109/tnnls.2013.2276044.

[9] NAYYERI M., YAZDI H.S., MASKOOKI A., ROUHANI M. Universal approximation by using the correntropy objective function. *IEEE Trans. Neural Netw. Learn. Syst.* 2018, 29(9), pp. 4515–4521, doi: 10.1109/tnnls.2017.2753725.

[10] ANDRAS P. High-dimensional function approximation with neural networks for large volumes of data. *IEEE Trans. Neural Netw. Learn. Syst.* 2018, 29(2), pp. 500–508, doi: 10.1109/tnnls.2017.2651985.

[11] MIRCHANDANI G., CAO W. On hidden nodes for neural nets. *IEEE Trans. Circ. Syst.* 1989, 36, pp. 661–664, doi: 10.1109/31.31313.

[12] HUANG S.C., HUANG Y.F. Bounds on the number of hidden neurons in multilayer perceptrons. *IEEE Trans. Neural Networks.* 1991, 2, pp. 47–55, doi: 10.1109/iscas.1990.112518.

[13] SARTORI M.A., ANTSAKLIS P.J. A simple method to derive bounds on the size and to train multilayer neural networks. *IEEE Trans. Neural Networks.* 1991, 2, pp. 467–471, doi: 10.1109/72.88168.

[14] TAMURA S., TATEISHI M. Capabilities of a four-Layered feedforward neural network: Four layers versus three. *IEEE Trans. Neural Networks.* 1997, 8(2), pp. 251–255, doi: 10.1109/72.557662.

[15] HUANG G.B. Learning capability and storage capacity of two-hidden- layer feedforward networks. *IEEE Trans. Neural Networks.* 2003, 14(2), pp. 274–281, doi: 10.1109/tnn.2003.809401.

[16] HUANG G.B., BABRI H.A. Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions. *IEEE Trans. Neural Networks.* 1998, 9(1), pp. 224–229, doi: 10.1109/72.655045.

[17] HUANG G.B., ZHU Q.Y., SIEW C.K. Extreme learning machine: Theory and applications. *Neurocomputing.* 2006, 70, pp. 489–501, doi: 10.1016/j.neucom.2005.12.126.

[18] FENG G., HUANG G.B., LIN Q., GAY R. Error minimized extreme learning machine with growth of hidden nodes and incremental learning. *IEEE Trans. Neural Networks.* 2009, 20(8), pp. 1352–1357, doi: 10.1109/tnn.2009.2024147.

[19] FU A.M., WANG X.Z., HE Y.L., WANG L.S. A study on residence error of training an extreme learning machine and its application to evolutionary algorithms. *Neurocomputing.* 2014, 146, pp. 75–82, doi: 10.1016/j.neucom.2014.04.067.

[20] KASSANI P.H., KIM E. Pseudoinverse matrix decomposition based incremental extreme learning machine with growth of hidden nodes. *International Journal of Fuzzy Logic and Intelligent Systems.* 2016, 16(2), pp. 125–130, doi: 10.5391/ijfis.2016.16.2.125.

[21] LU Y., HAN J.H., GAO J. Research on the minimal upper bound of the number of hidden nodes in binary neural networks. *Patt. Recog. Artif. Intell.* 2000, 13, pp. 254–257.

[22] LU Y., YANG J., WANG Q., HUANG Z.J. The upper bound of the minimal number of hidden neurons for the parity problem in binary neural networks. *Science China Information Sciences.* 2012, 55(7), pp. 1579–1587, doi: 10.1007/s11432-011-4405-6.

[23] ZHANG Z.Z., MA X.M., YANG Y.X. Bounds on the number of hidden neurons in three-layer binary neural networks. *Neural Networks.* 2003, 16(7), pp. 995–1002, doi: 10.1016/s0893-6080(03)00006-6.

[24] ARAI M. Bounds on the number of hidden units in binary-valued three-layer neural networks. *Neural Networks.* 1993, 6 (6), pp. 855–860, doi: 10.1016/s0893-6080(05)80130-3.

[25] TSAIH R., WAN Y. A guide for the upper bound on the number of continuous-valued hidden nodes of a feed-forward network. In: *International Conference on Artificial Neural Networks (ICANN 2009).* Lecture Notes in Computer Science, 2009, 5768, pp. 658–667, doi: 10.1007/978-3-642-04274-4_68.

[26] CAU G.W., FANG Z., CHEN Y.F. Estimating the number of hidden nodes of the single-hidden-layer feedforward neural networks. In: *15th International Conference on Computational Intelligence and Security (CIS 2019)*, Macao, China. IEEE, 2019, pp. 172–176, doi: 10.1109/cis.2019.00044.

[27] CUI R.Y., HONG B.R. Constructing a hidden layer for three-layered feedforward neural networks. *J. Comput. Res. Develop.* 2004, 41, pp. 524–530, doi: 10.1007/BF02873091.

[28] FUNG H.K., LI L.K. Minimal feedforward parity networks using threshold gates. *Neural. Comput.* 2001, 13, pp. 319–326, doi: 10.1162/089976601300014556.

[29] WANG H.C. On the compactness and the minimization. *Taiwanese Journal of Mathematics.* 2002, 6(4), pp. 441–464, doi: 10.11650/twjm/1500407470.

[30] ZHONG C.K., FAN X.L., CHEN W. *Introduction to nonlinear functional analysis (in Chinese).* Lanzhou: Lanzhou University Press, 2004.

Appendix

Proof of Proposition 2. The statement can be verified by an induction.

When $j = 1$, from Definition 1, we have the equations

$$\mathbf{W}^{(1)}\mathbf{u}_1 + \mathbf{b}_1 = \mathbf{d}_{1,1}, \ldots, \mathbf{W}^{(1)}\mathbf{u}_k + \mathbf{b}_1 = \mathbf{d}_{1,k},$$

which can be rewritten as

$$\begin{bmatrix}
\mathbf{u}_1^{\mathrm{T}} & & & & 1 & & & \\
 & \mathbf{u}_1^{\mathrm{T}} & & & & 1 & & \\
 & & \ddots & & & & \ddots & \\
 & & & \mathbf{u}_1^{\mathrm{T}} & & & & 1 \\
\mathbf{u}_2^{\mathrm{T}} & & & & 1 & & & \\
 & \mathbf{u}_2^{\mathrm{T}} & & & & 1 & & \\
 & & \ddots & & & & \ddots & \\
 & & & \mathbf{u}_2^{\mathrm{T}} & & & & 1 \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
\mathbf{u}_k^{\mathrm{T}} & & & & 1 & & & \\
 & \mathbf{u}_k^{\mathrm{T}} & & & & 1 & & \\
 & & \ddots & & & & \ddots & \\
 & & & \mathbf{u}_k^{\mathrm{T}} & & & & 1
\end{bmatrix} * \begin{bmatrix}
w_{11}^{(1)} \\ w_{12}^{(1)} \\ \vdots \\ w_{1e_0}^{(1)} \\ w_{21}^{(1)} \\ \vdots \\ w_{2e_0}^{(1)} \\ \vdots \\ w_{e_11}^{(1)} \\ \vdots \\ w_{e_1e_0}^{(1)} \\ b_{11} \\ \vdots \\ b_{1e_1}
\end{bmatrix} = \begin{bmatrix}
\mathbf{d}_{1,1} \\ \mathbf{d}_{1,2} \\ \vdots \\ \mathbf{d}_{1,k}
\end{bmatrix}. \tag{1}$$

It follows that the 1-layer pre-output combination vector $(\mathbf{d}_{1,1}{}^{\mathrm{T}}, \mathbf{d}_{1,2}{}^{\mathrm{T}}, \dots, \mathbf{d}_{1,k}{}^{\mathrm{T}})^{\mathrm{T}}$ is a linear combination of the column vectors of the first matrix in (1). Note that the column vectors in the first matrix in (1) consist the e_1-degree expansion vector list of $\overline{\mathrm{Ro}(\mathbf{u}_1, \dots, \mathbf{u}_k)}$. Hence, $\mathrm{OP}_1^{[e]} = \mathrm{Bon}^{e_1}(\mathrm{Span}(\overline{\mathrm{Ro}(\mathbf{u}_1, \dots, \mathbf{u}_k)}))$. Since $\delta_{e_0}(\mathrm{Ro}(\mathbf{u}_1, \dots, \mathbf{u}_k)) = \mathrm{Ro}(\mathbf{u}_1, \dots, \mathbf{u}_k)$ and f_0 is an identity mapping, we have $\mathrm{OP}_1^{[e]} = \mathrm{Bon}^{e_1}(\Psi_{f_0}^{\delta_{e_0}}(\mathrm{Ro}(\mathbf{u}_1, \dots, \mathbf{u}_k)))$. The statement holds for $j = 1$.

Analogously, from the equations $\mathbf{W}^{(j+1)}F(\mathbf{d}_{j,i}) + \mathbf{b}_{j+1} = \mathbf{d}_{j+1,i}, i = 1, \dots, k$, we have

$$\mathrm{OP}_{j+1}^{[e]} = \bigcup_{(\mathbf{d}_{j,1}{}^{\mathrm{T}}, \dots, \mathbf{d}_{j,k}{}^{\mathrm{T}})^{\mathrm{T}} \in \mathrm{OP}_j^{[e]}} \mathrm{Bon}^{e_{j+1}}(\mathrm{Span}(\overline{\mathrm{Ro}(f_j(\mathbf{d}_{j,1}), \dots, f_j(\mathbf{d}_{j,k}))})). \tag{2}$$

Assume that Proposition 2 holds for j. Now we prove that it holds for $j+1$. First, it will be verified that if $\mathbf{x} \in \mathrm{OP}_{j+1}^{[e]}$, then

$$\mathbf{x} \in \bigcup_{\delta_i \in \Delta_i} \mathrm{Bon}^{e_{j+1}} \left(\prod_{i=0}^{j} \Psi_{f_i}^{\delta_{e_i}} (\mathrm{Ro}(\mathbf{u}_1, \dots, \mathbf{u}_k)) \right).$$

By Lemma 2, it suffices to prove that each ro-column-vector of $\mathbf{x}$ belongs to

$$\bigcup_{\delta_{e_i} \in \Delta_{e_i}} \prod_{i=0}^{j} \Psi_{f_i}^{\delta_{e_i}} (\mathrm{Ro}(\mathbf{u}_1, \dots, \mathbf{u}_k)).$$

Since $\mathbf{x} \in \mathrm{OP}_{j+1}^{[e]}$, there exists $(\tilde{\mathbf{d}}_{j,1}^{\mathrm{T}}, \dots, \tilde{\mathbf{d}}_{j,k}^{\mathrm{T}})^{\mathrm{T}} \in \mathrm{OP}_{j+1}^{[e]}$, such that each ro-column-vector of $\mathbf{x}$ belongs to $\mathrm{Span}(\overline{\mathrm{Ro}(f_j(\tilde{\mathbf{d}}_{i,1}), \dots, f_j(\tilde{\mathbf{d}}_{i,k}))})$, according to Equation (2) and Lemma 2. Since Proposition 2 holds for j, there exist $\tilde{\delta}_{e_1} \in \Delta_{e_1}, \dots,$

$\tilde{\delta}_{e_{j-1}} \in \Delta_{e_{j-1}}$, such that

$$\left(\tilde{\mathbf{d}}_{j,1}^{\mathrm{T}}, \dots, \tilde{\mathbf{d}}_{j,k}^{\mathrm{T}}\right)^{\mathrm{T}} \in \mathrm{Bon}^{e_j}\left(\prod_{i=0}^{j-1} \Psi_{f_i}^{\tilde{\delta}_{e_i}}(\mathrm{Ro}(\mathbf{u}_1, \dots, \mathbf{u}_k))\right).$$

Each column vector in $\mathrm{Ro}(\tilde{\mathbf{d}}_{j,1}, \dots, \tilde{\mathbf{d}}_{j,k})$ belongs to $\prod\limits_{i=0}^{j-1} \Psi_{f_i}^{\tilde{\delta}_{e_i}}(\mathrm{Ro}(\mathbf{u}_1, \dots, \mathbf{u}_k))$ due to Lemma 2. Therefore, each column vector in $\mathrm{Ro}(f_j(\tilde{\mathbf{d}}_{j,1}), \dots, f_j(\tilde{\mathbf{d}}_{j,k}))$ belongs to

$$f_j\left(\prod_{i=0}^{j-1} \Psi_{f_i}^{\tilde{\delta}_{e_i}}(\mathrm{Ro}(\mathbf{u}_1, \dots, \mathbf{u}_k))\right).$$

Notice that $\mathrm{Ro}(f_j(\tilde{\mathbf{d}}_{j,1}), \dots, f_j(\tilde{\mathbf{d}}_{j,k}))$ has e_j column vectors, therefore, there exists $\tilde{\delta}_{e_j} \in \Delta_{e_j}$ such that

$$\mathrm{Ro}(f_j(\tilde{\mathbf{d}}_{j,1}), \dots, f_j(\tilde{\mathbf{d}}_{j,k})) \subseteq \tilde{\delta}_{e_j} f_j\left(\prod_{i=0}^{j-1} \Psi_{f_i}^{\tilde{\delta}_{e_i}}(\mathrm{Ro}(\mathbf{u}_1, \dots, \mathbf{u}_k))\right),$$

implying

$$\mathrm{Span}(\overline{\mathrm{Ro}(f_j(\tilde{\mathbf{d}}_{j,1}), \dots, f_j(\tilde{\mathbf{d}}_{j,k}))}) \subseteq \prod_{i=0}^{j} \Psi_{f_i}^{\tilde{\delta}_{e_i}}(\mathrm{Ro}(\mathbf{u}_1, \dots, \mathbf{u}_k)).$$

Consequently, each ro-column-vector of $\mathbf{x}$ belongs to $\prod\limits_{i=0}^{j} \Psi_{f_i}^{\tilde{\delta}_{e_i}}(\mathrm{Ro}(\mathbf{u}_1, \dots, \mathbf{u}_k))$.

Secondly, if $\mathbf{x} \in \bigcup\limits_{\delta_i \in \Delta_i} \mathrm{Bon}^{e_{j+1}}\left(\prod\limits_{i=0}^{j} \Psi_{f_i}^{\delta_{e_i}}(\mathrm{Ro}(\mathbf{u}_1, \dots, \mathbf{u}_k))\right)$, then $\mathbf{x} \in \mathrm{OP}_{j+1}^{[e]}$, which is proved as follows.

Since $\mathbf{x} \in \bigcup\limits_{\delta_i \in \Delta_i} \mathrm{Bon}^{e_{j+1}}\left(\prod\limits_{i=0}^{j} \Psi_{f_i}^{\delta_{e_i}}(\mathrm{Ro}(\mathbf{u}_1, \dots, \mathbf{u}_k))\right)$, there exist $\tilde{\delta}_{e_1} \in \Delta_{e_1}, \dots,$ $\tilde{\delta}_{e_j} \in \Delta_{e_j}$, such that each ro-column-vector of $\mathbf{x}$ belongs to $\prod\limits_{i=0}^{j} \Psi_{f_i}^{\tilde{\delta}_{e_i}}(\mathrm{Ro}(\mathbf{u}_1, \dots, \mathbf{u}_k))$ by Lemma 2.

Let $\tilde{\delta}_{e_j} f_j\left(\prod\limits_{i=0}^{j-1} \Psi_{f_i}^{\tilde{\delta}_{e_i}}(\mathrm{Ro}(\mathbf{u}_1, \dots, \mathbf{u}_k))\right)$ be $\{f_j(\mathbf{m}_1), \dots, f_j(\mathbf{m}_{e_j})\}$, with $\mathbf{m}_1, \dots,$ $\mathbf{m}_{e_j} \in \prod\limits_{i=0}^{j-1} \Psi_{f_i}^{\tilde{\delta}_{e_i}}(\mathrm{Ro}(\mathbf{u}_1, \dots, \mathbf{u}_k))$. Let $\{\mathbf{y}_1, \dots, \mathbf{y}_k\}$ be the column vectors in $\mathrm{Ro}(\mathbf{m}_1, \dots,$ $\mathbf{m}_{e_j})$. Then by Lemma 1, the column vectors in $\mathrm{Ro}(\mathbf{y}_1, \dots, \mathbf{y}_k)$ are $\{\mathbf{m}_1, \dots, \mathbf{m}_{e_j}\}$, and belong to $\prod\limits_{i=0}^{j-1} \Psi_{f_i}^{\tilde{\delta}_{e_i}}(\mathrm{Ro}(\mathbf{u}_1, \dots, \mathbf{u}_k))$. Consequently, each ro-column-vector of $\mathbf{x}$ belongs to $\mathrm{Span}(\overline{f_j(\mathbf{m}_1), \dots, f_j(\mathbf{m}_{e_j})})$=$\mathrm{Span}(\overline{\mathrm{Ro}(f_j(\mathbf{y}_1), \dots, f_j(\mathbf{y}_k))})$. Besides, $(\mathbf{y}_1{}^{\mathrm{T}}, \dots, \mathbf{y}_k{}^{\mathrm{T}})^{\mathrm{T}} \in \mathrm{Bon}^{e_j}\left(\prod\limits_{i=0}^{j-1} \Psi_{f_i}^{\tilde{\delta}_{e_i}}(\mathrm{Ro}(\mathbf{u}_1 \dots, \mathbf{u}_k))\right) \subseteq \mathrm{OP}_j^{[e]}$ by Lemma 2. Thus,

$$\mathbf{x} \in \mathrm{Bon}^{e_{j+1}}(\mathrm{Span}(\overline{\mathrm{Ro}(f_j(\mathbf{y}_1), \dots, f_j(\mathbf{y}_k))})) \subseteq \mathrm{OP}_{j+1}^{[e]}$$

by Lemma 2.