Universal Approximation Theorem for Tessarine-Valued Neural Networks

Conference Paper *in* Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC) · November 2021

DOI: 10.5753/eniac.2021.18256

3 authors, including:

Wington L. Vital
University of Campinas
6 PUBLICATIONS **5** CITATIONS

SEE PROFILE

Marcos Eduardo Valle
University of Campinas
140 PUBLICATIONS **1,284** CITATIONS

SEE PROFILE

Universal Approximation Theorem for Tessarine-Valued Neural Networks*

Rafael A. F. Carniello, Wington L. Vital, and Marcos Eduardo Valle

[1]Institute of Mathematics, Statistics, and Scientific Computing
University of Campinas (UNICAMP), Campinas, Brazil

{w265003, r204844}@dac.unicamp.br and valle@ime.unicamp.br

Abstract. *The universal approximation theorem ensures that any continuous real-valued function defined on a compact subset can be approximated with arbitrary precision by a single hidden layer neural network. In this paper, we show that the universal approximation theorem also holds for tessarine-valued neural networks. Precisely, any continuous tessarine-valued function can be approximated with arbitrary precision by a single hidden layer tessarine-valued neural network with split activation functions in the hidden layer. A simple numerical example, confirming the theoretical result and revealing the superior performance of a tessarine-valued neural network over a real-valued model for interpolating a vector-valued function, is presented in the paper.*

1. Introduction

Despite the many neural network architectures, this article focuses on single hidden layer neural networks, also known as multilayer perceptron (MLP). Once fixed the synaptic weights of the network, each input pattern is mapped into one unique (and always the same) output through the propagation of neural activation. As a consequence, MLPs are naturally able to handle static problems, such as regression and classification.

The universal approximation theorem is an existential theoretical result that justifies why neural networks can be used in practice. It states that any continuous real-valued function defined on a compact set can be approximated with arbitrary precision by a single hidden layer neural network. The universal approximation theorem for neural networks with a single hidden layer with sigmoid activation functions has been proved 1989 by Cybenko [Cybenko 1989]. Then, it was generalized for neural networks with arbitrary bounded and nonconstant activation functions by Hornik in 1991 [Hornik 1991]. Recently, many researchers addressed the approximation capabilities of neural networks, including deep and shallow models based on piece-wise linear activations functions such as the widely used rectified linear unit (ReLU) activation function [Hanin and Sellke 2017, Lu et al. 2017, Yarotsky 2017, Petersen and Voigtlaender 2018].

In contrast to the widely used real-valued MLP model, a quaternion-valued feedforward network was developed by Arena et al. in the late 1990s [Arena et al. 1997]. Quaternions are four-dimensional hypercomplex numbers widely used to describe spatial rotations [Arena et al. 1998]. Like real-valued models, quaternion-valued neural networks have been effectively applied for classification and regression problems

*This work was supported in part by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Table 1. Tessarines multiplication table

$\times$	$\boldsymbol{i}$	$\boldsymbol{j}$	$\boldsymbol{k}$
$\boldsymbol{i}$	-1	$\boldsymbol{k}$	$-\boldsymbol{j}$
$\boldsymbol{j}$	$\boldsymbol{k}$	1	$\boldsymbol{i}$
$\boldsymbol{k}$	$-\boldsymbol{j}$	$\boldsymbol{i}$	-1

[Parcollet et al. 2020]. Moreover, quaternion-valued networks usually reduce the number of parameters for processing multi-dimensional data. Like the real-valued network, any continuous quaternion-valued function on a compact can be approximated by a single hidden layer quaternion-valued MLP (Q-MLP) network whose hidden units are equipped with the so-called split activation function [Arena et al. 1997].

Despite the quaternions developed by Hamilton in 1843, James Cockle presented the tessarine numbers in a series of articles published in 1848. Like the quaternions, tessarines are also four-dimensional hypercomplex numbers. On the one hand, some tessarines do not have a multiplicative inverse. On the other hand, in contrast to quaternion's algebra, the product of tessarines is commutative. The commutativity of tessarine's product may be advantageous for the design of neural networks because synaptic weights and neuron inputs can interchange their roles. Thus, this paper uses tessarine algebra to define a new class of neural networks called tessarine-valued multilayer perceptron (T-MLP). Moreover, we point out that the universal approximation theorem holds for T-MLP networks. In other words, we show that any continuous tessarine-valued function defined on a compact can be approximated by a single hidden layer T-MLP network with arbitrary precision.

The paper is organized as follows: Next section presents the tessarine algebra and its properties. Section 3 contains some theoretical results and auxiliary definitions for the main theorem. The universal approximation theorem for T-MLP is proved in Section 4. A simple numerical example comparing real-valued and tessarine-valued MLP networks is given in Section 5. The paper finishes with some concluding remarks in Section 6.

2. Tessarines Algebra

Tessarines are hypercomplex four-dimensional hypercomplex numbers represented as follows

$$t = t_0 + t_1\boldsymbol{i} + t_2\boldsymbol{j} + t_3\boldsymbol{k}$$

where $t_0, t_1, t_2, t_3 \in \mathbb{R}$ and $\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$ are the hypercomplex units. Throughout this paper, the set of all tessarines is denoted by $\mathbb{T}$. The tessarine hypercomplex units $\boldsymbol{i}$, $\boldsymbol{j}$, and $\boldsymbol{k}$ satisfy the multiplication Table 1. In contrast to the multiplication table of quaternions, Table 1 is symmetric. As a consequence, the product of tessarines is commutative.

Given two tessarines $t = t_0 + t_1\boldsymbol{i} + t_2\boldsymbol{j} + t_3\boldsymbol{k}$ and $w = w_0 + w_1\boldsymbol{i} + w_2\boldsymbol{j} + w_3\boldsymbol{k}$, their sum is

$$t + w = t_0 + w_0 + (t_1 + w_1)\boldsymbol{i} + (t_2 + w_2)\boldsymbol{j} + (t_3 + w_3)\boldsymbol{k}. \tag{1}$$

The product of t and w is established using distributivity and the multiplication table

resulting in:

$$\begin{aligned} tw &= (t_0 + t_1\boldsymbol{i} + t_2\boldsymbol{j} + t_3\boldsymbol{k})(w_0 + w_1\boldsymbol{i} + w_2 j + w_3\boldsymbol{k}) \\ &= (t_0w_0 - t_1w_1 + t_2w_2 - t_3w_3) + (t_0w_1 + t_1w_0 + t_2w_3 + t_3w_2)\boldsymbol{i} \\ &\quad + (t_0w_2 - t_1w_3 + t_2w_0 + t_3w_3)\boldsymbol{j} + (t_0w_3 + t_1w_2 + t_2w_1 + t_3w_0)\boldsymbol{k} \end{aligned} \tag{2}$$

A tessarine $t = t_0 + t_1\boldsymbol{i} + t_2\boldsymbol{j} + t_3\boldsymbol{k}$ can also be written as

$$t = R(t) + I(t)\boldsymbol{i} + J(t)\boldsymbol{j} + K(t)\boldsymbol{k},$$

where $R, I, J, K : \mathbb{T} \to \mathbb{R}$ are the functions defined by

$$R(t) = t_0, \quad I(t) = t_1, \quad J(t) = t_2, \quad \text{and} \quad K(t) = t_3, \tag{3}$$

for all $t = t_0 + t_1\boldsymbol{i} + t_2\boldsymbol{j} + t_3\boldsymbol{k} \in \mathbb{T}$. In this case, $t_0 = R(t)$ is the real part of t while $t_1 = I(t)$, $t_2 = J(t)$, and $t_3 = K(t)$ are the imaginary parts of t. The absolute value of a tessarine $t = t_0 + t_1\boldsymbol{i} + t_2\boldsymbol{j} + t_3\boldsymbol{k}$ is defined by

$$|t| = \sqrt{t_0^2 + t_1^2 + t_2^2 + t_3^2}. \tag{4}$$

The nth Cartesian product of $\mathbb{T}$ is

$$\mathbb{T}^n = \{\boldsymbol{t} = (t_1, \cdots, t_n) : t_i \in \mathbb{T}\}. \tag{5}$$

The set $\mathbb{T}^n$ is interpreted as a vector space. Like in real-valued linear algebra, given two tessarine-valued vectors $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{T}^n$ and $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathbb{T}^n$, we define

$$\boldsymbol{x}^T \cdot \boldsymbol{y} = x_1y_1 + x_2y_2 + \ldots + x_ny_n. \tag{6}$$

Note that the inner product given by (6) satisfies the following properties for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{T}^n$ and $\lambda \in \mathbb{R}$:

1. $\boldsymbol{x}^T \cdot (\boldsymbol{y} + \boldsymbol{z}) = \boldsymbol{x}^T \cdot \boldsymbol{y} + \boldsymbol{x}^T \cdot \boldsymbol{z}$.
2. $\boldsymbol{x}^T \cdot (\lambda\boldsymbol{y}) = \lambda\boldsymbol{x}^T \cdot \boldsymbol{y}$.
3. $\boldsymbol{x}^T \cdot \boldsymbol{y} = \boldsymbol{y}^T \cdot \boldsymbol{x}$.

Finally, we would like to point out that the set of all tessarine numbers can be identified with $\mathbb{R}^4$ by means of the isomorphism $\psi : \mathbb{T} \to \mathbb{R}^4$ defined by

$$\psi(t) = (t_0, t_1, t_2, t_4), \quad \forall t = t_0 + t_1\boldsymbol{i} + t_2\boldsymbol{j} + t_3\boldsymbol{k}. \tag{7}$$

Similarly, the Cartesian product $\mathbb{T}^n$ can be identified with $\mathbb{R}^{4n}$ by applying ψ in a component-wise manner, that is,

$$(t_1, \ldots, t_n) \in \mathbb{T}^n \overset{\psi}{\longleftrightarrow} (\psi(t_1), \ldots, \psi(t_n)) \in \mathbb{R}^{4n}. \tag{8}$$

3. Basic Concepts on Approximation Theory

Let us begin by presenting a class of activation functions used for proving the universal approximation theorem for neural networks [Cybenko 1989]:

Definition 1. *A function $\phi : \mathbb{R} \to \mathbb{R}$ is called discriminatory if the only signed Borel measure μ on the n-dimensional unit hypercube $I_n = [0, 1]^n$ such that*

$$\int_{I_n} \phi(\boldsymbol{y}^T \cdot \boldsymbol{x} + \theta) d\mu(\boldsymbol{x}) = 0, \forall \boldsymbol{y} \in \mathbb{R}^n \text{ and } \forall \theta \in \mathbb{R}, \tag{9}$$

is the zero measure, that is, $\mu = 0$.

The continuous sigmoid function $\sigma(x) = 1/(1 + e^{-x})$ is an example of a discriminatory function. The linear rectified function (ReLU), defined by $\texttt{ReLU}(x) = \max\{0, x\}$ for all $x \in \mathbb{R}$ and widely used for designing neural networks nowadays, is also a discriminatory function [Ferreira Guilhoto]. The following presents the universal approximation theorem for neural networks with discriminatory functions in the hidden layer.

Theorem 1 (Universal Approximation Theorem [Cybenko 1989])**.** *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a continuous discriminatory function. Then, the class of all real-valued neural networks defined by*

$$\mathcal{H}_{\mathbb{R}} = \left\{ \mathcal{N}_{\mathbb{R}}(\boldsymbol{x}) = \sum_{i=1}^{M} \alpha_i \phi(\boldsymbol{y}_i^T \cdot \boldsymbol{x} + \theta_i), \forall \boldsymbol{x} \in \mathbb{R}^n \right\}, \tag{10}$$

is dense in the class $\mathcal{C}(K)$ of all real-valued continuous functions on a compact $K \subseteq \mathbb{R}^n$. In other words, given $f_{\mathbb{R}} : K \to \mathbb{R}$ and $\epsilon > 0$, there exist $M > 0$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M) \in \mathbb{R}^M$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M) \in \mathbb{R}^M$, and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M \in \mathbb{R}^n$ such that the single hidden-layer neural network given by

$$\mathcal{N}_{\mathbb{R}}(\boldsymbol{x}) = \sum_{i=1}^{M} \alpha_i \phi(\boldsymbol{y}_i^T \cdot \boldsymbol{x} + \theta_i), \quad \forall \boldsymbol{x} \in \mathbb{R}^n, \tag{11}$$

satisfies

$$|f_{\mathbb{R}}(\boldsymbol{x}) - \mathcal{N}_{\mathbb{R}}(\boldsymbol{x})| < \epsilon, \quad \forall \boldsymbol{x} \in K. \tag{12}$$

Because $\ell_i(\boldsymbol{x}) = \boldsymbol{y}^T \cdot \boldsymbol{x}$ is a linear functional, the universal approximation theorem states that a function $f_{\mathbb{R}} : K \to \mathbb{R}$ can be approximated by a single hidden-layer neural network given by

$$\mathcal{N}_{\mathbb{R}}(\boldsymbol{x}) = \sum_{i=1}^{M} \alpha_i \phi(\ell_i(\boldsymbol{x}) + \theta_i), \quad \forall \boldsymbol{x} \in \mathbb{R}^n, \tag{13}$$

where $\ell_1, \ell_2, \ldots, \ell_M : \mathbb{R}^n \to \mathbb{R}$ are linear functions.

4. Universal Approximation Theorem for Tessarine-Valued Neural Networks

Briefly, an hypercomplex-valued neural network is obtained by replacing the real-valued inputs, outputs, and parameters by hypercomplex-valued entities. A quaternion-valued feedforward neural network has been developed by [Arena et al. 1997]. Besides introducing quaternion-valued networks, Arena et al. proved that single hidden-layer neural networks are universal approximators when the hidden units have discriminatory

split functions. In the following, we derive a similar result for tessarine-valued MLP networks.

Consider a tessarine-valued activation function $\phi : \mathbb{T} \to \mathbb{T}$. Given tessarine-valued vectors $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M) \in \mathbb{T}^M$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M) \in \mathbb{T}^M$, and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M \in \mathbb{T}^n$, a tessarine-valeud MLP network (T-MLP) is defined by

$$\mathcal{N}_{\mathbb{T}}(\boldsymbol{x}) = \sum_{i=1}^{M} \alpha_i \phi(\boldsymbol{y}_i^T \cdot \boldsymbol{x} + \theta_i), \quad \forall \boldsymbol{x} \in \mathbb{T}^n. \tag{14}$$

Note the algebraic similarity between the real-valued and the tessarine-valued neural networks given by (11) and (14), respectively. In contrast to the real-valued neural network, the T-MLP model defines an application $\mathcal{N}_{\mathbb{T}} : \mathbb{T}^n \to \mathbb{T}$. Thus, like quaternion-valued neural networks, the T-MLP model process four-dimensional information as a single entity. An up-to-date review on quaternion-valued neural networks and their applications can be found in [Parcollet et al. 2020]. We believe that many applications of quaterion-valued neural networks can also be addressed by tessarine-valued models. In the following, we show that the universal approximation theorem also holds for the T-MLP networks. To this end, let us present some concepts and Lemma 1.

The following lemma, which is analogous to Lemma 3.1 of [Arena et al. 1997], plays a key role in the proof of the universal approximation theorem for the T-MLP networks:

Lemma 1. *Given a linear function $\ell : \mathbb{T}^n \to \mathbb{R}$, there are uniquely determined tessarine-valued vectors $\boldsymbol{y}_R, \boldsymbol{y}_I, \boldsymbol{y}_J, \boldsymbol{y}_K \in \mathbb{T}^n$ such that*

$$\ell(\boldsymbol{t}) = R(\boldsymbol{y}_R^T \cdot \boldsymbol{t}) = I(\boldsymbol{y}_I^T \cdot \boldsymbol{t}) = J(\boldsymbol{y}_J^T \cdot \boldsymbol{t}) = K(\boldsymbol{y}_K^T \cdot \boldsymbol{t}), \quad \forall \boldsymbol{t} \in \mathbb{T}^n.$$

Proof. Let us prove the result for the representation $\ell(\boldsymbol{t}) = R(\boldsymbol{y}_R^T \cdot \boldsymbol{t})$. The other three representations are analogous. Consider the application

$$L : \mathbb{T}^n \to \{\ell : \ell \text{ is a linear function from } \mathbb{T}^n \text{ to } \mathbb{R}\},$$

defined by setting $L(\boldsymbol{y}) = \ell$ with $\ell(\boldsymbol{t}) = R(\boldsymbol{y}^T \cdot \boldsymbol{t})$ for every $\boldsymbol{t} \in \mathbb{T}^n$. Because the linear product and the projection R are linear, the application L is also linear. Moreover, from the isomorphism between $\mathbb{T}$ and $\mathbb{R}^4$, we conclude that $dim(\mathbb{T}^n) = 4n$. Similarly,

$$dim(\{\ell : \ell \text{ is a linear function from } \mathbb{T}^n \text{ to } \mathbb{R}\}) = dim(\mathbb{T}^n) dim(\mathbb{R}) = 4n.$$

Therefore, the linear application L maps spaces of the same dimension. In the following, we show that $Ker(L) = \{0\}$ which imples that L is a one-to-one correspondence.

Consider $\boldsymbol{y} = (y_1, \ldots, y_n) \in Ker(L)$. In this case, we have $L(\boldsymbol{y}) = 0$ or, equivalently, $R(\boldsymbol{y}^T \cdot \boldsymbol{t}) = 0$ for every $\boldsymbol{t} \in \mathbb{T}^n$. In particular, considering $\boldsymbol{t} \in \{\boldsymbol{e}_j, \boldsymbol{i}\boldsymbol{e}_j, \boldsymbol{j}\boldsymbol{e}_j, \boldsymbol{k}\boldsymbol{e}_j\}$, where $\boldsymbol{e}_j$ denotes the jth canonical basis of $\mathbb{R}^n$ for $j = 1, \ldots, n$, we conclude that

$$\begin{aligned}
0 &= R\big(\boldsymbol{y}^T \cdot \boldsymbol{e}_j\big) = R(y_j), \\
0 &= R\big(\boldsymbol{y}^T \cdot (\boldsymbol{i}\boldsymbol{e}_j)\big) = -I(y_j), \\
0 &= R\big(\boldsymbol{y}^T \cdot (\boldsymbol{j}\boldsymbol{e}_j)\big) = J(y_j), \\
0 &= R\big(\boldsymbol{y}^T \cdot (\boldsymbol{k}\boldsymbol{e}_j)\big) = -K(y_j).
\end{aligned}$$

Therefore, $y_j = R(y_j)+\boldsymbol{i}I(y_j)+\boldsymbol{j}J(y_j)+\boldsymbol{k}K(y_j) = 0$ for any $j = 1, \ldots, n$. Equivalently, $\boldsymbol{y} = 0$ which implies that $Ker(L) = \{0\}$. Concluding, the mapping L is a one-to-one correspondence. Hence, given a linear function $\ell : \mathbb{T}^n \to \mathbb{R}$ one defines $\boldsymbol{y}_R = L^{-1}(\ell)$.

□

In the following we present the main contribution of this paper: the universal approximation theorem for T-MLP networks with split activation functions. A real-valued function $\phi : \mathbb{R} \to \mathbb{R}$ yields a tessarine-valued function $\phi_{\mathbb{T}} : \mathbb{T} \to \mathbb{T}$ defined as follows for all $t = t_0 + t_1\boldsymbol{i} + t_2\boldsymbol{j} + t_3\boldsymbol{k} \in \mathbb{T}$:

$$\phi_{\mathbb{T}}(t) = \phi(t_0) + \phi(t_1)\boldsymbol{i} + \phi(t_2)\boldsymbol{j} + \phi(t_3)\boldsymbol{k}. \tag{15}$$

In this case, we say that $\phi_{\mathbb{T}} : \mathbb{T} \to \mathbb{T}$ is a split function derived from $\phi : \mathbb{R} \to \mathbb{R}$. Note that a split function can be alternatively written as follows using the composition of ϕ and the functions $R, I, J, K : \mathbb{T} \to \mathbb{R}$ given by (3):

$$\phi_{\mathbb{T}}(t) = (\phi \circ R)(t) + (\phi \circ I)(t)\boldsymbol{i} + (\phi \circ J)(t)\boldsymbol{j} + (\phi \circ K)(t)\boldsymbol{k}. \quad \forall t \in \mathbb{T}. \tag{16}$$

Theorem 2. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a continuous discriminatory function such that $\lim_{\lambda\to-\infty} \phi(\lambda) = 0$ and $\phi_{\mathbb{T}} : \mathbb{T} \to \mathbb{T}$ be the split function derived from ϕ by means of* (15)*. Then, the class of all tessarine-valued neural networks defined by*

$$\mathcal{H}_{\mathbb{T}} = \left\{ \mathcal{N}_{\mathbb{T}}(\boldsymbol{t}) = \sum_{i=1}^{M} \alpha_i \phi_{\mathbb{T}}(\boldsymbol{y}_i^T \cdot \boldsymbol{t} + \theta_i), \forall \boldsymbol{t} \in \mathbb{T}^n \right\}, \tag{17}$$

is dense in the class $\mathcal{C}(K)$ of all tessarine-valued continuous functions on a compact $K \subseteq \mathbb{T}^n$. In other words, given $f_{\mathbb{T}} : K \to \mathbb{T}$ and $\epsilon > 0$, there exist $M > 0$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M) \in \mathbb{T}^M$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M) \in \mathbb{T}^M$, and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M \in \mathbb{T}^n$ such that the single hidden-layer neural network given by

$$\mathcal{N}_{\mathbb{T}}(\boldsymbol{t}) = \sum_{i=1}^{M} \alpha_i \phi_{\mathbb{T}}(\boldsymbol{y}_i^T \cdot \boldsymbol{t} + \theta_i), \quad \forall \boldsymbol{x} \in \mathbb{T}^n, \tag{18}$$

satisfies

$$|f_{\mathbb{T}}(\boldsymbol{t}) - \mathcal{N}_{\mathbb{R}}(\boldsymbol{t})| < \epsilon, \quad \forall \boldsymbol{t} \in K, \tag{19}$$

where $|\cdot|$ denotes the absolute value of a tessarine.

Proof. First of all, from (3), the continuous tessarine-valued function $f_{\mathbb{T}} : \mathbb{T}^n \to \mathbb{T}$ can be written as follows for all $\boldsymbol{t} \in \mathbb{T}$:

$$f_{\mathbb{T}}(\boldsymbol{t}) = (R \circ f_{\mathbb{T}})(\boldsymbol{t}) + \boldsymbol{i}(I \circ f_{\mathbb{T}})(\boldsymbol{t}) + \boldsymbol{j}(J \circ f_{\mathbb{T}})(\boldsymbol{t}) + \boldsymbol{k}(K \circ f_{\mathbb{T}})(\boldsymbol{t}). \tag{20}$$

Let us now apply the universal approximation theorem for real-valued neural networks on the real and imaginary parts of $f_{\mathbb{T}}$ by identifying a tessarine-valued vector in $\mathbb{T}^n$ with a real-valued vector in $\mathbb{R}^{4n}$ by means of the isomorphism (8). Precisely, given $\epsilon > 0$, from

Theorem 1, there exist an integer $M > 0$, real-valued vectors $\alpha = (\alpha_1, \ldots, \alpha_M) \in \mathbb{R}^M$ and $\theta = (\theta_1, \ldots, \theta_M) \in \mathbb{R}^M$, and linear functions $\ell_1, \ldots, \ell_M : \mathbb{T}^n \to \mathbb{R}$ such that

$$\left| (R \circ f_{\mathbb{T}})(\boldsymbol{t}) - \sum_{i=1}^{M} \alpha_i \phi\Big(\ell_i(\boldsymbol{t}) + \theta_i\Big) \right| < \frac{\epsilon}{2}, \quad \forall \boldsymbol{t} \in K. \tag{21}$$

From Lemma (1), there exist tessarine-valued vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M$ such that

$$\left| (R \circ f_{\mathbb{T}})(\boldsymbol{t}) - \sum_{i=1}^{M} \alpha_i \phi\Big(R\big(\boldsymbol{y}_i^T \cdot \boldsymbol{t}\big) + \theta_i\Big) \right| < \frac{\epsilon}{2}, \quad \forall \boldsymbol{t} \in K. \tag{22}$$

Now, we define

$$\theta_i(\lambda) = \theta_i + \lambda \boldsymbol{i} + \lambda \boldsymbol{j} + \lambda \boldsymbol{k} \in \mathbb{T}, \quad \forall i = 1, \ldots, M. \tag{23}$$

Then, because $\lim_{\lambda \to -\infty} \phi(\lambda) = 0$, we have

$$\lim_{\lambda \to -\infty} \phi_{\mathbb{T}}\Big(\boldsymbol{y}_i^T \cdot \boldsymbol{t} + \theta_i(\lambda)\Big) = \phi\Big(R\big(\boldsymbol{y}_i^T \cdot \boldsymbol{t}\big) + \theta_i\Big) + 0\boldsymbol{i} + 0\boldsymbol{j} + 0\boldsymbol{k}. \tag{24}$$

As a consequence, there exists λ such that

$$\left| \sum_{i=1}^{M} \alpha_i \phi_{\mathbb{T}}\Big(\boldsymbol{y}_i^T \cdot \boldsymbol{t} + \theta_i(\lambda)\Big) - \sum_{i=1}^{M} \alpha_i \phi\Big(R\big(\boldsymbol{y}_i^T \cdot \boldsymbol{t}\big) + \theta_i\Big) \right| < \frac{\epsilon}{2}. \tag{25}$$

Using the triangle inequality and the inequalities (22) and (25), we obtain

$$\left| (R \circ f_{\mathbb{T}})(\boldsymbol{t}) - \sum_{i=1}^{M} \alpha_i \phi_{\mathbb{T}}\Big(\boldsymbol{y}_i^T \cdot \boldsymbol{t} + \theta_i(\lambda)\Big) \right| \leq \left| (R \circ f_{\mathbb{T}})(\boldsymbol{t}) - \sum_{i=1}^{M} \alpha_i \phi\Big(R\big(\boldsymbol{y}_i^T \cdot \boldsymbol{t}\big) + \theta_i\Big) \right| \tag{26}$$

$$+ \left| \sum_{i=1}^{M} \alpha_i \phi\Big(R\big(\boldsymbol{y}_i^T \cdot \boldsymbol{t}\big) + \theta_i\Big) - \sum_{i=1}^{M} \alpha_i \phi_{\mathbb{T}}\Big(\boldsymbol{y}_i^T \cdot \boldsymbol{t} + \theta_i(\lambda)\Big) \right| \leq \epsilon \tag{27}$$

In words, there exist tessarine-valued vectors $\alpha = (\alpha_1, \ldots, \alpha_M)$, $\theta(\lambda) = (\theta_1(\lambda), \ldots, \theta_M(\lambda)) \in \mathbb{T}^M$, and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M \in \mathbb{T}^n$ such that the real-part of $f_{\mathbb{T}}(\boldsymbol{t})$ can be approximated with arbitrary precision by $\sum_{i=1}^{M} \alpha_i \phi_{\mathbb{T}}(\boldsymbol{y}_i^T \cdot \boldsymbol{t} + \theta_i(\lambda))$ for some λ. In a similar fashion, we can show that

$$\left| (I \circ f_{\mathbb{T}})(\boldsymbol{t}) - \sum_{i=1}^{M} \alpha_i \phi_{\mathbb{T}}\Big(\boldsymbol{y}_i^T \cdot \boldsymbol{t} + \theta_i(\lambda)\Big) \right| \leq \epsilon, \tag{28}$$

where

$$\theta_i(\lambda) = \lambda + \theta_i \boldsymbol{i} + \lambda \boldsymbol{j} + \lambda \boldsymbol{k}, \quad \forall i = 1, \ldots, M, \tag{29}$$

and $\alpha = (\alpha_1, \ldots, \alpha_M)$, $\theta = (\theta_1, \ldots, \theta_M)$, and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M \in \mathbb{T}^n$ are obtained by applying Theorem 1 to approximate $(I \circ f_{\mathbb{T}})$ and Lemma 1. Finally, the approximation of the other two imaginary parts of $f_{\mathbb{T}}$ are derived similarly. □

Note that the sigmoid function $\sigma(x) = 1/(1 + e^{-x})$ and the linear rectified function $\texttt{ReLU}(x) = \max\{0, x\}$ are both continuous discriminatory functions such that $\lim_{\lambda \to -\infty} \phi(\lambda) = 0$. Therefore, Theorem 2 holds, in particular, for tessarine-valued networks equipped with these two activation functions.

5. Numerical Example

Let us now provide a simple example to illustrate Theorem 2.

Consider the non-linear tessarine-valued function $f_{\mathbb{T}} : K \subseteq \mathbb{T} \to \mathbb{T}$ given by

$$f_{\mathbb{T}}(t) = pt + tq + t^2, \quad \forall t \in K, \tag{30}$$

where

$$p = 1.5 - 1.3\boldsymbol{i} + 1.2\boldsymbol{j} + 0.5\boldsymbol{k} \quad \text{and} \quad q = 0.5 + 0.3\boldsymbol{i} + 0.2\boldsymbol{j} - \boldsymbol{k}, \tag{31}$$

and $K \subseteq \mathbb{T}$ is the compact set defined by

$$K = \{t = t_0 + t_1\boldsymbol{i} + t_2\boldsymbol{j} + t_3\boldsymbol{k} \in \mathbb{T} : 0 \leq t_0, t_1, t_2, t_3 \leq 1\}. \tag{32}$$

We would like to point out that the function $f_{\mathbb{T}}$ given by (30) is analogous to the quaternion-valued function considered by Arena et al. to illustrate the universal approximation theorem for quaternion-valued single-hidden layer neural network [Arena et al. 1994].

To illustrate Theorem 2, we considered a tessarine-valued network with two hidden neurons equipped with the linear rectified function. The parameters of the T-MLP network are tessarines y_1, y_2, θ_1, θ_2, α_1, and α_2, resulting a total 24 ($= 6 \times 4$) real-valued parameters. Moreover, the tessarine-valued synaptic weights have been adjusted by minimizing the mean squared error (MSE) using Adam optimizer with 1000 epochs on a training set determined as follows with $N = 400$:

$$\left\{\left(t_i, f_{\mathbb{T}}(t_i)\right) : t_i = r_{i0} + r_{i1}\boldsymbol{i} + r_{i2}\boldsymbol{j} + r_{i3}\boldsymbol{k}, r_{i0}, r_{i1}, r_{i2}, r_{i3} \sim U[-1, +1], i = 1, \ldots, N\right\}. \tag{33}$$

Then, we evaluated the performance of the tessarine-valued network using a test set obtained from (33) with $N = 100$. The real and the imaginary parts of both analytical and approximated outputs by the testing samples are shown in Figures 1 - 4. Also, the first row of Table (2) shows the MSE yielded by the real and imaginary parts of the T-MLP on the training set. Note that the tessarine-valued network yielded errors in order of magnitude less than 10^{-7}. Therefore, the T-MLP network was able to approximate the tessarine-valued function $f_{\mathbb{T}}$ given by (30).

For comparison purposes, we also approximated the function $f_{\mathbb{T}}$ by a real-valued network. Precisely, using the isomorphism $\psi : \mathbb{T} \to \mathbb{R}^4$ given by (7), we approximated $\psi \circ f_{\mathbb{T}} \circ \psi^{-1}$ by a real-valued network with $4 - 6 - 4$ architecture equipped with the linear rectified activation function. The real-valued MLP network has 54 adjustable parameters. The synaptic weights and bias of the real-valued network have been adjusted analogously to the tessarine-valued model. The second row of Table 2 shows the MSE product by the real-valued network for the real and imaginary parts of the output on the test set. As expected, the real-valued network can also approximate the function $f_{\mathbb{T}}$ but with MSEs of orders of 10^{-4} and 10^{-5}.

where:

$$\overline{h} = [1.5, -1.3, 1.2, 0.5]^T$$
$$\overline{f} = [[0.5, 0.3, 0.2, -1]^T$$

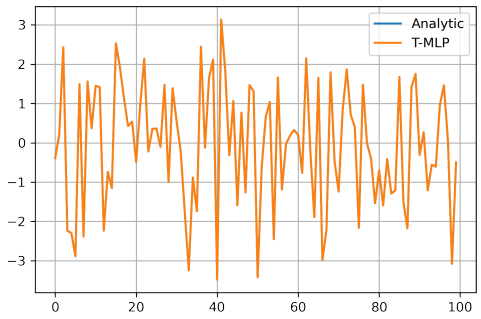

Figure 1. Comparison between T-MLP output and the target for the real part.

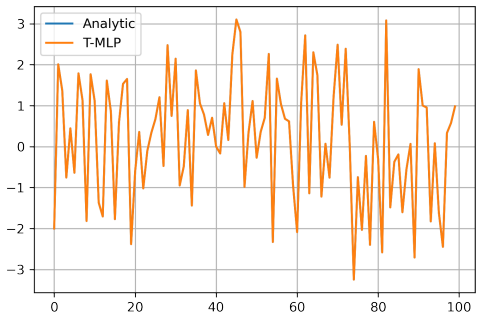

Figure 2. Comparison between T-MLP output and the target for the first imaginary component.

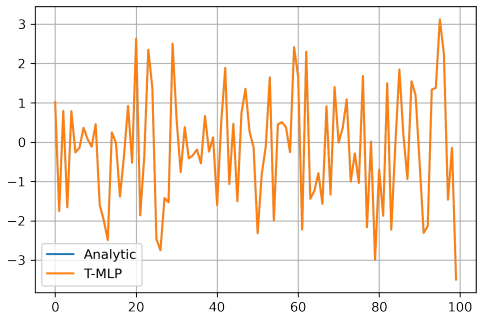

Figure 3. Comparison between T-MLP output and the target for the second imaginary component.

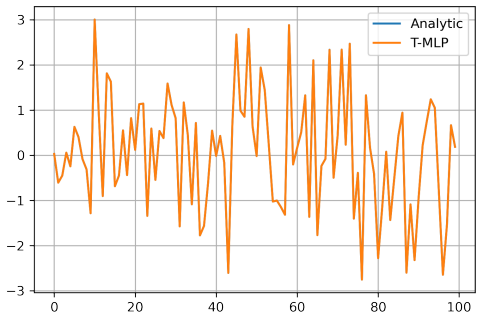

Figure 4. Comparison between T-MLP output and the target for the third imaginary component.

with

$$\{\overline{q} \in \mathbb{R} : -1 \leq \overline{q} \leq 1\}$$

in which $\overline{q}$ is a tessarine.

For the proposed interpolation it was generated a uniform sample with 500 values of $\overline{q}$, such that, 100 where utilized for training, and the other 400 for evaluating the network capability of generalization. The architecture used for the tessarine model was $1 - 2 - 1$, with ReLU as activation function in the hidden layer. Finally, the optimizer chosen was Adam with loss function MSE, such that 1000 epochs were computed. The results obtained for the function with tessarine values in the four components of the hypercomplex numbers were:

Comparing the mean squared error between the TMLP and the reference model for each component:

By the charts it is possible to notice that, the TMLP, for each one of it's components, was able to approximate the reference function. In fact, the 1000 epochs of training represent the power of interpolation of that kind of net, which was capable to generate an overfit. In practical terms, it is expected to monitor the loss function to optimize the

MLP	r	i	j	k
Tessarine	2.7×10^{-7}	2.3×10^{-7}	6.8×10^{-8}	4.8×10^{-7}
Real	2.6×10^{-4}	1.5×10^{-4}	4.6×10^{-5}	4.5×10^{-5}

Table 2. Comparison of the mean squared error from the TMLP and MLP output for each component.

generalization of the model, for example, using an early stop regularization.

Moreover, by the table, we can compare the accuracy between the TMLP and the canonical MLP with real values. This, kept the same architecture characteristics, that is, $4-2-4$, given by the isomorphism with tessarines in $\mathbb{R}^4$. So, for both the real component and the imaginary ones, the TMLP obtained an error in order of magnitude of 10^{-4}. That error order was only obtained by the real MLP in the second imaginary component, with the others in the order of 10^{-3}.

6. Concluding Remarks

In this paper, we addressed the universal approximation capability of tessarine-valued feedforward single-hidden layer neural networks. Broadly speaking, the tessarine-valued networks are obtained by replacing the real-valued inputs, outputs, and parameters by four-dimensional hypercomplex numbers whose multiplication satisfy Table (1). Like the real-valued networks, any tessarine-valued function can be approximated with arbitrary precision by a tessarine-valued multilayer perceptron (T-MLP) on a compact set. In particular, T-MLP networks with sigmoid or linear rectified activation functions enjoy the universal approximation capability. We finished the paper with a single numerical example to illustrate the approximation capability of the tessarine-valued networks.

In the future, we plan to investigate further tessarine-valued neural networks. In particular, we plan to investigate their applicability for solving classification and regression problems.

Therefore, the tessarine based neural network is a powerful interpolator. Comparatively, it showed a superior performance in reference with the real valued MLP for approximating tassarine functions. That proof, in theoretical terms, represents the fundamental mechanism of new category of nets - TMLP. In practical terms, opens new doors for exploring applications in a diverse range of scientific problem with hypercomplex neural networks.

References

[Arena et al. 1997] Arena, P., Fortuna, L., Muscato, G., and Xibilia, M. G. (1997). Multilayer perceptrons to approximate quaternion valued functions. *Neural Networks*, 10(2):335–342.

[Arena et al. 1998] Arena, P., Fortuna, L., Muscato, G., and Xibilia, M. G. (1998). Quaternion algebra. In *Neural Networks in Multidimensional Domains*, volume 234 of *Lecture Notes in Control and Information Sciences*, pages 43–47. Springer London.

[Arena et al. 1994] Arena, P., Fortuna, L., Occhipinti, L., and Xibilia, M. G. (1994). Neural networks for quaternion-valued function approximation. *Proceedings - IEEE International Symposium on Circuits and Systems*, 6:307–310.

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

[Ferreira Guilhoto] Ferreira Guilhoto, L. An Overview Of Artificial Neural Networks for Mathematicians.

[Hanin and Sellke 2017] Hanin, B. and Sellke, M. (2017). Approximating Continuous Functions by ReLU Nets of Minimal Width.

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

[Lu et al. 2017] Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. (2017). The Expressive Power of Neural Networks: A View from the Width. *Advances in Neural Information Processing Systems*, 2017-December:6232–6240.

[Parcollet et al. 2020] Parcollet, T., Morchid, M., and Linarès, G. (2020). A survey of quaternion neural networks. *Artificial Intelligence Review*, 53(4):2957–2982.

[Petersen and Voigtlaender 2018] Petersen, P. and Voigtlaender, F. (2018). Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108:296–330.

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