LETTER Communicated by Simon Haykin

Approximation by Fully Complex Multilayer Perceptrons

Taehwan Kim
tkim@mitre.org
MITRE Corporation, McLean, Virginia 22102 U.S.A.

Tülay Adalı
adali@umbc.edu
Department of Computer Science and Electrical Engineering, University of Maryland Baltimore County, Baltimore, Maryland 21250 U.S.A.

We investigate the approximation ability of a multilayer perceptron (MLP) network when it is extended to the complex domain. The main challenge for processing complex data with neural networks has been the lack of bounded and analytic complex nonlinear activation functions in the complex domain, as stated by Liouville's theorem. To avoid the conflict between the boundedness and the analyticity of a nonlinear complex function in the complex domain, a number of ad hoc MLPs that include using two real-valued MLPs, one processing the real part and the other processing the imaginary part, have been traditionally employed. However, since nonanalytic functions do not meet the Cauchy-Riemann conditions, they render themselves into degenerative backpropagation algorithms that compromise the efficiency of nonlinear approximation and learning in the complex vector field. A number of elementary transcendental functions (ETFs) derivable from the entire exponential function e^z that are analytic are defined as fully complex activation functions and are shown to provide a parsimonious structure for processing data in the complex domain and address most of the shortcomings of the traditional approach. The introduction of ETFs, however, raises a new question in the approximation capability of this fully complex MLP. In this letter, three proofs of the approximation capability of the fully complex MLP are provided based on the characteristics of singularity among ETFs. First, the fully complex MLPs with continuous ETFs over a compact set in the complex vector field are shown to be the universal approximator of any continuous complex mappings. The complex universal approximation theorem extends to bounded measurable ETFs possessing a removable singularity. Finally, it is shown that the output of complex MLPs using ETFs with isolated and essential singularities uniformly converges to any nonlinear mapping in the deleted annulus of singularity nearest to the origin.

1 Introduction

The main challenge of extending the multilayer perceptron (MLP) paradigm to process complex-valued data has been the lack of bounded and analytic complex nonlinear activation functions in the complex plane $\mathbb{C}$ that need to be incorporated in the hidden (and sometimes output) layers of a neural network. As stated by Liouville's theorem, a bounded entire function must be a constant in $\mathbb{C}$, where an entire function is defined as analytic (i.e., differentiable at every point $z \in \mathbb{C}$; Silverman, 1975). Therefore, we cannot find an analytic complex nonlinear activation function that is bounded everywhere in $\mathbb{C}$. The fact that many real-valued mathematical objects are best understood when they are considered as embedded in the complex plane (e.g., the Fourier transform) provides a basic motivation to develop a neural network structure that uses well-defined analytic nonlinear activation functions (Clarke, 1990). However, due to Liouville's theorem, the complex MLP has been subjected to the common view that it has to live with a compromised employment of nonanalytic but bounded nonlinear activation functions for the stability of the system (Georgiou & Koutsougeras, 1992; You & Hong, 1998; Mandic & Chambers, 2001). Hence, to process the complex data, a split complex approach is typically adopted (Leung & Haykin, 1991; Benvenuto, Marchesi, Piazza, & Uncini, 1991). Another approach has been to use two independent real-valued MLPs, dealing with the real and the imaginary part separately or with the magnitude and the phase. However, because of the cumbersome representation of separated real and imaginary (or magnitude and phase) components, all of these approaches are inefficient in representing complex-valued functions. Also, the nonanalytic but bounded ad hoc complex functions cannot provide truthful complex gradients required for the error backpropagation process. Therefore, the nonanalytic activation functions render themselves into degenerative forms for the backpropagation of error that compromise the efficiency of complex number representation and learning paradigm suitable for further optimization. Consequently, the traditional complex MLPs are often unable to achieve robust and resilient tracking and pattern matching performances in noisy and time-varying signal conditions.

In Kim and Adalı (2000), a subset of elementary transcendental functions (ETFs) derivable from the exponential function $f(z) = e^z$, which are entire since $f(z) = f'(z) = e^z$ in $\mathbb{C}$, is proposed as activation functions for nonlinear processing of complex data. The fully complex MLP is shown to provide better performance compared to the traditional MLPs in system identification (Kim, 2002), equalization for nonlinear satellite channel (Kim & Adalı, 2001, 2002), and nonlinear prediction and cancellation of scintillation examples (Kim, 2002; Kim & Hegarty, 2002).

In this letter, *fully complex activation function* is used to refer to one of the 10 ETFs shown in section 2.2 that is entire and meets Cauchy-Riemann conditions. The fully complex MLP employs one of these ETFs and takes

advantage of compact joint real and imaginary input feedforward and error backpropagation through well-defined complex error gradients.

In this letter, we study the approximation properties of these activation functions when they are used in a feedforward structure and present three key results that explain their approximation ability. The ETFs we consider are entire and either bounded in a bounded domain of interest or bounded almost everywhere (a.e.), that is, unbounded only on a set of points having zero measures in $\mathbb{C}$. Since these functions are entire, they are also conformal and provide well-defined derivatives while still allowing the complex MLP to converge with probability 1, which is sufficient in most practical applications of MLPs. Also, among these ETFs, circular and hyperbolic functions such as $\tan z$ and $\tanh z$ are a special form of bilinear (i.e., Möbius) transforms and lend themselves to the fixed-point analysis discussed in Mandic and Chambers (2001). Note, however, that inverse functions such as $\arcsin z$ and $\operatorname{arcsinh} z$ are not bilinear transforms, but perform very well in certain learning tasks (Kim, 2002) because of their elegant symmetric and squashing magnitude responses as discussed in section 2.2.

The organization of the rest of the article is as follows. The traditional complex-valued MLPs and error backpropagation algorithms that rely on the nonanalytic but bounded nonlinear activation functions are summarized in section 2. In section 3, three categories of singularity that ETFs inherently possess are presented, and their relationships to the approximation capabilities of fully complex MLP are established. Similar to its well-known real-valued MLP counterpart, it is shown that fully complex MLPs using continuous and bounded measurable activation functions are universal approximators over a compact set in the n-dimensional complex vector field $\mathbb{C}^n$ by using the Stone-Weierstrass and the Riesz representation theorems. For measurable but unbounded elementary transcendental functions with isolated singularities and nonmeasurable functions with essential singularity, we show the uniform convergence to any complex number with an arbitrary accuracy within the annulus of a deleted singularity via Laurent series approximation. We include a short discussion on some of the applications for which complex nonlinear processing is particularly important and present a simple numerical example to highlight the efficiency of fully complex representation. Section 5 contains a summary and discussions on the further research.

2 Complex-Valued MLPs

2.1 Traditional Complex-Valued MLPs. This section gives a brief summary of previous work on complex-valued MLPs. (A more detailed discussion can be found in Kim, 2002.) It was Clarke (1990) who generalized the real-valued activation function into the complex-valued one. He suggested employing a complex tangent sigmoidal function $\tanh z = (e^z - e^{-z})/(e^z + e^{-z}), z \in \mathbb{C}$ and noted that this function is no longer bounded but in-

cludes singularities. He added that to keep the activation function analytic, one cannot avoid having unbounded functions because of Liouville's theorem. However, after this important observation was made, the expansion of neural networks in the complex domain followed a series of ad hoc approaches instead of further investigation of the unbounded nature of analytic activation functions. The first and the most commonly used ad hoc approach has been the employment of the pair of $\tanh x, x \in \mathbb{R}$—hence, the name *split complex activation function* (Leung & Haykin, 1991; Benvenuto et al., 1991). This approach has inherent limitations in representing a truthful gradient and, consequently, for developing the fully complex backpropagation algorithm because the derivatives of split-complex activation function cannot fully represent the true gradient unless the real and imaginary weight updates in the error backpropagation are completely independent. However, this is clearly impossible in a complex error backpropagation process. Equation 2.1 represents the split-complex activation function,

$$f(z) = f_R(\mathrm{Re}(z)) + if_I(\mathrm{Im}(z)), \tag{2.1}$$

where a pair of real-valued sigmoidal functions $f_R(x) = f_I(x) = \tanh x, x \in \mathbb{R}$ process the real and imaginary components of $z = W \cdot X + b$, that is, inner product of the complex input vector X and the complex weight vector W of the same dimension plus a complex bias. The split-complex backpropagation algorithm almost simultaneously developed in Leung and Haykin (1991) and Benvenuto et al. (1991) did employ complex arithmetic but was shown to be a degenerative and special form of fully complex backpropagation algorithm (Kim & Adalı, 2001, 2002).

A compromised approach to process real and imaginary components of complex signal jointly followed soon (Georgiou & Koutsougeras, 1992; Hirose, 1992). These authors proposed joint-nonlinear complex activation functions that process the real and imaginary components as shown in equations 2.2 and 2.3, respectively:

$$f(z) = z/(c + |z|/r) \tag{2.2}$$

$$f\left(s \cdot \exp[i\beta]\right) = \tanh(s/m)\exp[i\beta]. \tag{2.3}$$

Here, c and r are real positive constants, and m is a constant that is inversely related to the gradient of the absolute function $|f|$ along the radius direction around the origin of the complex coordinate for $z = s \cdot \exp[i\beta]$. However, these functions are still not analytic and also preserve the phase. The inability to provide accurate nonlinear phase response poses a significant disadvantage for these functions in signal processing applications, as shown in section 4.

Recently, it has been shown that the analytic nonlinear function $\tanh z$ can successfully be used in a fully complex MLP structure where its superior performance over the traditional nonanalytic approach is demonstrated in

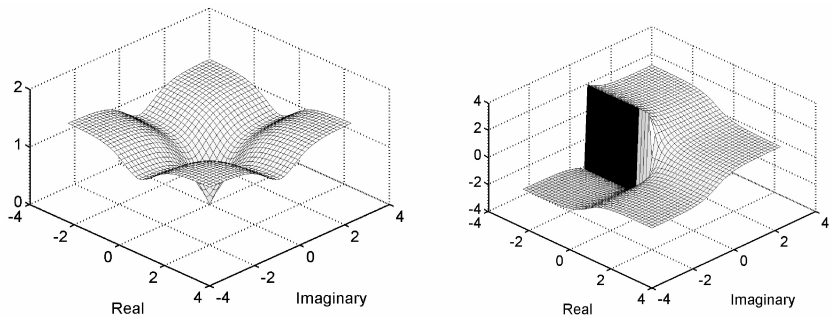

Figure 1: Split tanh z. (Left) Magnitude. (Right) Phase.

a nonlinear channel equalization example (Kim & Adalı, 2000). The article also develops the fully complex error backpropagation algorithm that takes advantage of the well-defined true complex gradient satisfying the Cauchy-Riemann equations. The split and joint-complex backpropagation algorithms are also shown to be special and degenerative cases of the fully complex backpropagation algorithm and that the fully complex backpropagation algorithm is a complex conjugate form of its real-valued counterpart (Kim & Adalı, 2001, 2002). Nine other fully complex nonlinear activation functions from the elementary transcendental function family are identified to be adequate complex squashing functions. The advantage of having truthful gradients in these functions over the pseudo-gradients generated from split- and joint-complex activation functions is demonstrated in various numerical examples (Kim & Adalı, 2001, 2002; Kim & Hegarty, 2002; Kim, 2002).

Figure 1 shows the magnitude and phase characteristics of the split tanh z function given in equation 2.1. Figure 2 shows the magnitude and phase of the joint-complex activation function proposed by Georgiou and Koutsougeras given in equation 2.2, followed by Figure 3 showing Hirose's joint-complex activation function given in equation 2.3. As observed in these figures, the nonlinear discrimination ability of the phase of these functions is limited, and the magnitude of split tanh z divides the domain into four quadrants.

2.2 Fully Complex Activation Functions. The following three classes of elementary transcendental functions have been identified as those that can provide adequate squashing-type nonlinear discrimination with well-defined first-order derivatives, and hence are candidates for developing a fully complex MLP (Kim & Adalı, 2001).

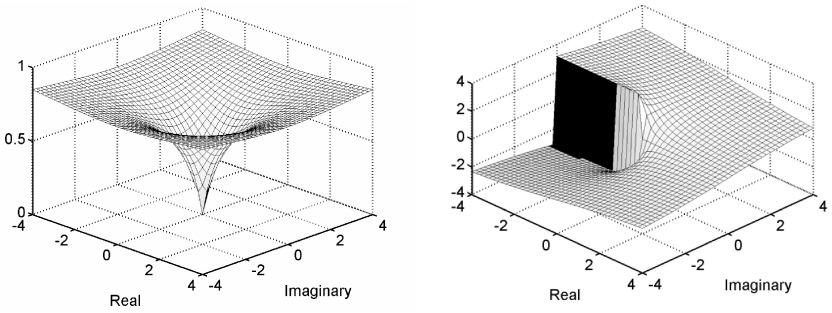

Figure 2: $f(z) = z/(c + |z|/r)$. (Left) Magnitude. (Right) Phase.

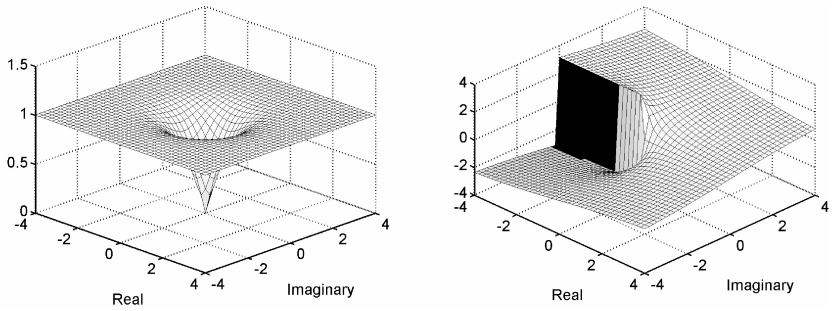

Figure 3: $\tanh(s/m)\exp[i\beta]$. (Left) Magnitude. (Right) Phase.

- Circular functions:

$$\tan z = \frac{e^{iz} - e^{-iz}}{i(e^{iz} + e^{-iz})}, \quad \frac{d}{dz}\tan z = \sec^2 z$$
$$\sin z = \frac{e^{iz} - e^{-iz}}{2i}, \quad \frac{d}{dz}\sin z = \cos z$$

- Inverse circular functions:

$$\arctan z = \int_0^z \frac{1}{1+t^2}, \quad \frac{d}{dz}\arctan z = \frac{1}{1+z^2}$$
$$\arcsin z = \int_0^z \frac{dt}{(1-t)^{1/2}}, \quad \frac{d}{dz}\arcsin z = (1-z^2)^{-1/2}$$
$$\arccos z = \int_0^z \frac{dt}{(1-t^2)^{1/2}}, \quad \frac{d}{dz}\arccos z = -(1-z^2)^{-1/2}$$

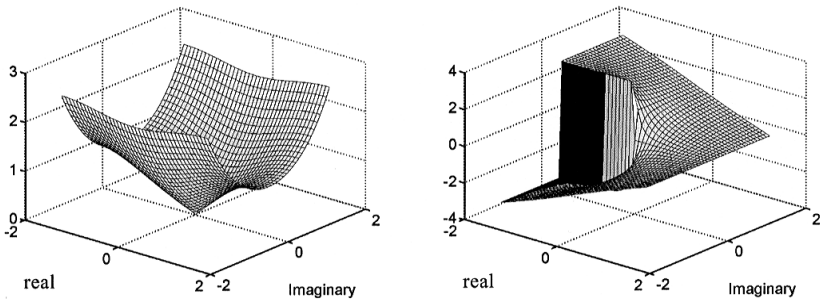

Figure 4: $\sin 2$. (Left) Magnitude. (Right) Phase.

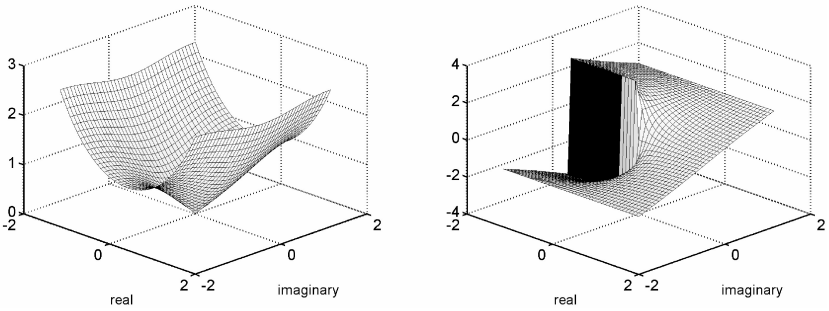

Figure 5: $\sinh z$. (Left) Magnitude. (Right) Phase.

- Hyperbolic functions:

$$\tanh z = \frac{\sinh z}{\cosh z} = \frac{e^z - e^{-z}}{e^z + e^{-z}}, \quad \frac{d}{dz}\tanh z = \operatorname{sech}^2 z$$

$$\sinh z = \frac{e^z - e^{-z}}{2}, \quad \frac{d}{dz}\sinh z = \cosh z$$

- Inverse hyperbolic functions:

$$\operatorname{arctanh} z = \int_0^z \frac{dt}{1-t^2}, \quad \frac{d}{dz}\operatorname{arctanh} z = (1-z^2)^{-1}$$

$$\operatorname{arcsinh} z = \int_0^z \frac{dt}{(1+t^2)^{1/2}}, \quad \frac{d}{dz}\operatorname{arcsinh} z = (1+z^2)^{-1}$$

Figures 4 through 12 show the magnitude and phase responses of these elementary transcendental functions.

Figures 4 and 5 (both left) show the magnitude of $\sin z$ and $\sinh z$ where the sine curve characteristics in the range $[-\pi/2, \pi/2]$ are observed along the

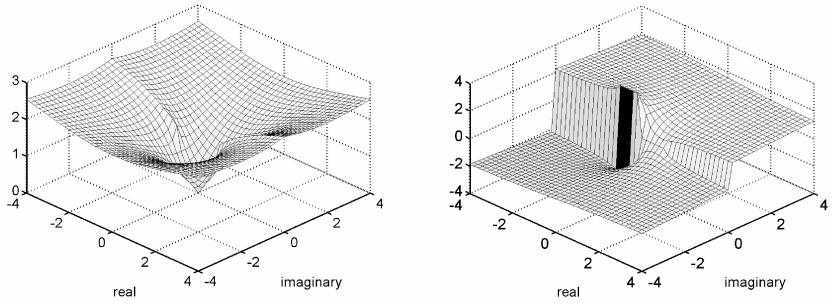

Figure 6: $\arcsin z$. (Left) Magnitude. (Right) Phase.

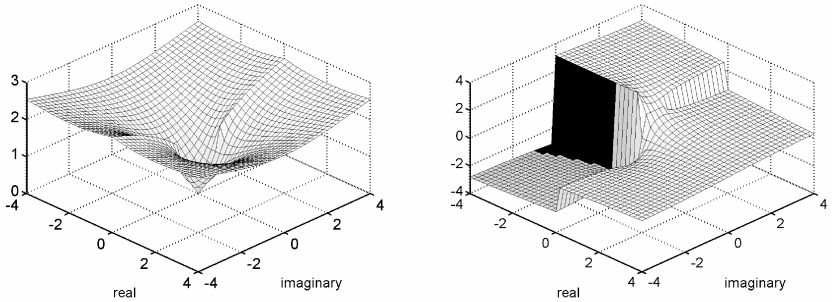

Figure 7: $\operatorname{arcsinh} z$. (Left) Magnitude. (Right) Phase.

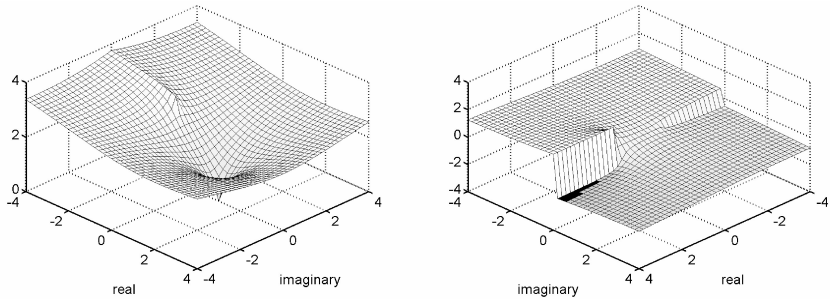

Figure 8: $\arccos z$. (Left) Magnitude. (Right) Phase.

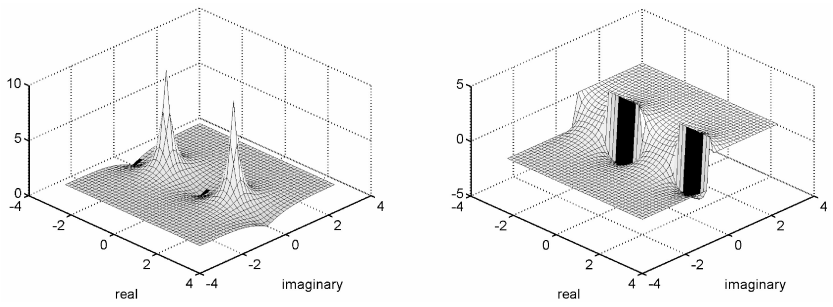

Figure 9: $\tan z$. (Left) Magnitude. (Right) Phase.

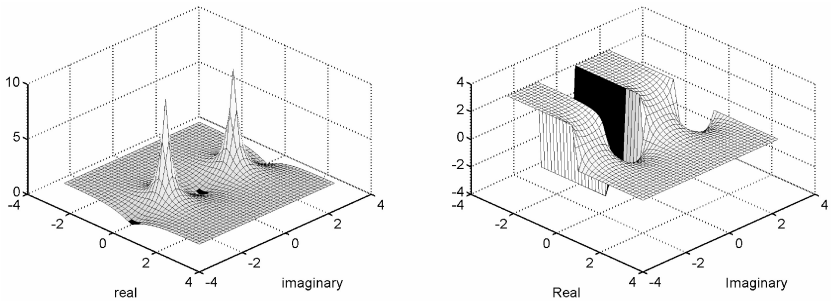

Figure 10: $\tanh z$. (Left) Magnitude. (Right) Phase.

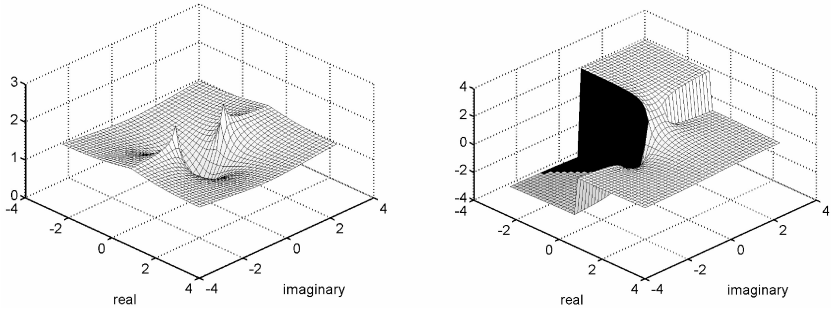

Figure 11: $\arctan z$. (Left) Magnitude. (Right) Phase.

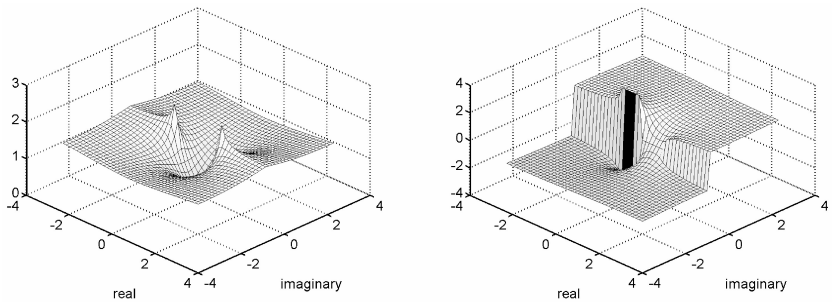

Figure 12: arctanh z. (Left) Magnitude. (Right) Phase.

real and imaginary axes, respectively. Unlike the other ETFs, those shown in Figures 6 through 12, $\sin z$ and $\sinh z$ functions exhibit continuous magnitude but are unbounded functions with convex behavior along the imaginary and real axis, respectively, as their domain expands from the origin. Therefore, these functions, when bounded in a radius of $\pi/2$ can be effective nonlinear approximators. Figures 6 and 7 (both left) show that $\arcsin z$ and $\operatorname{arcsinh} z$ functions are not continuous and therefore not analytic along a portion of the real or the imaginary axis. These discontinuities known as branch cuts are explained in section 3.2. However, these two are the the most radially symmetric functions in magnitude. As observed in figure 8 (left), $\arccos z$ function is asymmetric along the real axis and has a discontinuous zero point at $z = 1$. The split-tanh x function shown in Figure 1, on the other hand, does not possess similar smooth radial symmetry property because of the valleys it has along the real and imaginary axes. Note that these functions have decreasing rates of magnitude growth as they move away from the origin. Also note that unlike $\operatorname{arcsinh} z$ and $\arccos z$ functions that are unbounded, the split-tanh x function is bounded by $\sqrt{2}$ in magnitude as they are defined in terms of two real-valued bounded functions. Instead, when the domain is bounded, $\arcsin z$, $\operatorname{arcsinh} z$, and $\arccos z$ are naturally bounded while providing more discriminating nonlinearity than the split-tanh x function. It has been observed that the radial symmetricity as well as the nonlinear phase response of $\operatorname{arcsinh} z$ and $\arcsin z$ functions tend to provide efficient nonlinear approximation capability (Kim, 2002).

Unlike $\arcsin z$ and $\operatorname{arcsinh} z$ functions that have branch cut–type singularities, note the point-wise periodicity and singularity of $\tanh z$ at every $(1/2+n)\pi i$, $n \in \mathrm{N}$, in Figure 10 (left). Similarly, $\tan z$ is singular and periodic at every $(1/2+n)\pi$, $n \in \mathrm{N}$, as shown in Figure 9 (left). In contrast, Figures 11 and 12 (both left) show isolated singularities of $\arctan z$ and $\operatorname{arctanh} z$ at $\pm i$ and ± 1, respectively. Although our focus in this article is on the approximation using these ETFs as the activation functions, we note that these types of singular points and discontinuities at nonzero points do not pose a prob-

lem in training when the domain of interest is bounded within a circle of radius $\pi/2$. If the domain is larger and includes these irregular points and the initial random hidden layer weight radius is not small enough, then the training process tends to become more sensitive to the size of the learning rate and the radius of initial random weights (Kim, 2002).

3 Approximation by Fully Complex MLP

In this section, the theory of the approximation capability of fully complex MLP is developed. In this development, three types of approximation theorems emerge according to the three types of singularity that each of the elementary transcendental functions possesses (except the continuous $\sin z$ and $\sinh z$ functions that do not contain any).

The first two approximation results and the supporting domains of convergence for the first two classes of functions are very general and resemble the universal approximation theorem for the real-valued feedforward MLP that was shown almost concurrently by multiple authors in 1989 (Cybenko, 1989; Hornik, Stinchecombe, & White, 1989; Funahashi, 1989). The third approximation theorem for the fully complex MLP is unique and related to the power series approximation that can represent any complex number arbitrarily closely in the deleted neighborhood of a singularity. This approximation is uniform only in the analytic domain of convergence whose radius is defined by the closest singularity.

3.1 Background. The universal approximation theorem for real-valued MLPs shows the existence of arbitrarily accurate approximation of any continuous or Borel-measurable multivariate real mapping of one finite-dimensional space to another by using a finite linear combination of a fixed, univariate nonlinear activation function, provided sufficient hidden neuron units are available.

It is well known that Arnold and Kolmogorov refuted Hilbert's thirteenth conjecture in 1957 when Kolmogorov's superposition theorem solved Hilbert's arbitrary representation question of multivariate functions using bivariate functions (Cybenko, 1989; Hornik et al., 1989). Kolmogorov's superposition theorem showed that all continuous functions of n variables have an exact representation in terms of finite superposition and composition of a small number of functions of one or two variables.

However, the Kolmogorov theorem requires representation of different unknown nonlinear transformations for each continuous desired multivariate function, while specifying an exact upper limit to the number of intermediate units needed for the representation (Hornik et al., 1989). For MLP, the interest is in finite linear combinations involving the same univariate activation function for approximations as opposed to exact representations. Mathematically, the representation of the general function of an n-

dimensional real variable, $x \in \mathbb{R}^n$, by finite linear combinations of the form is expressed

$$G(x) = \sum_{k=1}^{N} \alpha_k \sigma(W_k^T x + \theta_k), \tag{3.1}$$

where $W_k \in \mathbb{R}^n$ and $\alpha_k, \theta_k \in \mathbb{R}$ are fixed. In the neural network approach, the nonlinear function σ is typically a sigmoidal function.

Cybenko noted that equation 3.1 could be seen as a special case of a generalized approximation by finite Fourier series. He demonstrated two mathematical tools to prove the completeness properties of such series, (or, equivalently, the uniform convergence of Cauchy sequences): the algebra of functions leading to the Stone-Weierstrass theorem (Rudin, 1991) and the translation-invariant subspaces leading to the Tauberian theorem. Hornik et al. (1989) also employed Stone-Weierstrass theorem with cosine transform approximation and then extended their universal approximation theorem to multi-output and multilayer feedforward MLP. Since Cybenko's proof for the single-output, single hidden-layer MLP has shown the fundamental universal approximation in a more compact manner, the approximation theorems in fully complex MLP that we present here for an arbitrary complex continuous and measurable mapping follow similar steps to Cybenko's derivation.

3.2 Classification of Activation Functions. In this section, we present the classification of complex activation functions based on their types of singularity. The first two classes divide the ETFs included in section 2.2 into two classes: a set of bounded squashing activation functions over the bounded domain and a set of activation functions with unbounded isolated singularities. We then proceed and introduce a third class, activation functions having essential singularity, and discuss the properties of each class to establish the universal approximation property. A single-valued function is said to have a singularity at a point if the function is not analytic, and therefore not continuous, at the point. In complex analysis, three types of singularities are known: removable, isolated, and essential (Silverman, 1975). Note that ETFs possessing singularities are classified as entire functions just like the complex exponential function $f(z) = f'(z) = e^z$ from which they all can be derived (Churchill & Brown, 1992; Needham 2000).

3.2.1 Removable Singularity Class. If $f(z)$ has single point singularity at z_0, the singularity is said to be removable if $\lim_{z \to z_0} f(z)$ exists. The inverse sine and cosines shown in Figures 6, 7, and 8 have removable singularities represented as ridges along the real or imaginary axis outside the unit circle, which is defined as the branch cuts. Without loss of generality, even if the complex sine and cosine functions shown in section 2.2 have periodic-

ity defined by their branch cuts (Silverman, 1975), the working domain is treated as connected and bounded by the branch cuts to ensure that they are single-valued functions. These branch cut discontinuities of inverse functions are related to the definition of the corresponding integral where the paths of integration must not cross the branch cuts. The inverse sine and cosine functions exhibit the unbounded but decreasing rate of magnitude growth as they move away from the origin. Therefore, once the domain is bounded and connected, the range of these functions is naturally bounded with squashing function characteristics.

On the other hand, $\sin z$ and $\sinh z$ functions are continuous and grow convex upward with an increasing rate of magnitude growth in parallel to real or imaginary axis, respectively, as shown in Figures 4 and 5. Also, note that $\sin z$ is equivalent to $\sin x$ along the real axis, while $\sinh z$ is equivalent to $\sin x$ along the imaginary axis. Therefore, the bounded squashing property for $\sin z$ and $\sinh z$ functions is available within a radius of $\pi/2$ from the origin.

Consequently, $\sin z$, $\sinh z$, $\arcsin z$, $\operatorname{arcsinh} z$, and $\arccos z$ functions (the sine family) can be classified as complex squashing functions within a bounded domain. Note that by using a scaling coefficient, the size of squashing domain can be controlled. The formal definition of a complex squashing function is given in the appendix.

3.2.2 Isolated Singularity Class. If $\lim_{z\to z_0} |f(z)| \to \infty$, while $f(z)$ is analytic in a deleted neighborhood of $z = z_0$, then the singularity is not removable but isolated. The isolated singularities are found in the tangent function family. The inverse tangent functions have nonperiodic isolated singularities: $\arctan z$ at $\pm i$ and $\operatorname{arctanh} z$ at ± 1. On the contrary, $\tanh z$ has isolated periodic singularities at every $(1/2+n)\pi i$, $n \in \mathbb{N}$, where $\lim_{z\to(1/2+n)\pi i^+} \tanh z = -\infty$ and $\lim_{z\to(1/2+n)\pi i^-} \tanh z = +\infty$. Similarly, $\tan z$ has periodic isolated singularities at every $(1/2 + n)\pi$.

3.2.3 Essential Singularity Class. If a singularity is neither removable nor isolated, then it is known as (isolated) essential singularity. A rare example can be found in $f(z) = e^{1/z}$. As shown in Figure 13 (left) in three dimensions and in Figure 13 (right) along the real axis, the function has singularity at 0 with multiple limiting values depending on the direction of approach: $\lim_{z\to 0^{\pm i}} e^{1/z} = 1$, $\lim_{z\to 0^-} e^{1/z} = 0$, and $\lim_{z\to 0^+} e^{1/z} = \infty$. For this reason, the essential singularity does not satisfy Cauchy-Riemman equations at the origin. However, the essential singularity has an intriguing property summarized in a powerful theorem known as the Big Picard theorem (or Picard's Great theorem) (Silverman, 1975). It states that in the neighborhood of an essential singularity, a function assumes each complex value, with one possible exception ($f(z) = e^z$), infinitely often. Therefore, if a priori knowledge on the working domain is available, a variation of $f(z) = e^{1/z}$, for example, $f(z) = 0.5(e^{1/(z-z_0)} + e^{1/(z+z_0)})$, can be used as an activation function

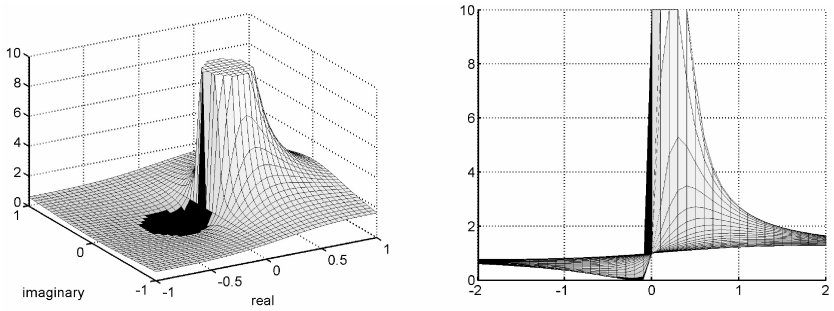

Figure 13: Essential singularity function $f(z) = e^{1/z}$. (Left) 3D magnitude. (Right) Real axis magnitude view.

by placing multiple shifted and scaled singularities to build a powerful approximation.

In the uniform approximation theorem using the essential singularity activation function, it is shown that for an activation function with an essential singularity centered at origin, the converging domain should exclude the essential singularity itself to form a deleted annulus.

3.3 Approximation by Fully Complex MLP. Since most of the ETFs are not continuous, except $\sin z$ and $\sinh z$, but have singularities, it is important to know whether they are complex meausurable functions, that is, the measure over the set of singularities is zero in the complex vector field. The definition of complex measurable space and (Borel) complex measurable functions are given in definition 4 in the appendix.

Note that the value infinity is admissible for a positive measure, but when a complex measure μ is being considered, defined on a σ-algebra $\Re$, it is understood that $\mu(E)$ is a complex number for every $E \in \Re$. The real measures form a subclass of the complex measures.

Now that the formal definitions of a complex measurable set and function are given, it is easy to verify that elementary transcendental functions with their well-defined inverse functions are complex measurable functions. For example, $f(z) = \tanh z$ is a measurable function since if X is the disc with radius $\pi/2$ at the origin, then Y is the whole complex plane $\mathbb{C}$ because of the isolated singularity at $\pi/2$ that yields to infinity. It is readily observable that the extreme value has inverse image $f^{-1}(\infty) = \pi/2 \in X$ and $f^{-1}(-\infty) = -\pi/2 \in X$ where $f^{-1}(z) = \operatorname{arctanh} z$, and vice versa.

Now let I_n denote the n-dimensional complex unit cube, $[0, 1]^n$ and $C(I_n)$ the space of all continuous complex functions on I_n. The space of finite, signed Borel measures on I_n is denoted by $M(I_n)$. Note that $C(I_n) \subset M(I_n)$. The major result of this section is to determine the conditions of approximation under which the single hidden-layer fully complex MLP outputs are

close to its target function in $M(I_n)$ (or $C(I_n)$) with respect to the supremum norm $\|\cdot\|$, where $\|z\| = (z\bar{z})^{1/2}$. For approximation between two complex functions f and g belonging to $C(I_n)$ or $M(I_n)$, their closeness is measured by a metric ρ. The general concept on the closeness of one class of functions to another class is described by the notion of denseness given in definition 5 in the appendix.

In words, a set of points S is dense in a set of points T if every neighborhood of each point t in T contains points of S. This can be further generalized to the complex continuous and Borel measurable functions in the form of uniform dense functional space (see definition 6).

Using these definitions, the first step for establishing the universal approximation theorem of a fully complex MLP is showing that its continuous output is uniformly dense on compacta in $C(I_n)$. This requires the following definitions to build an algebra on the outputs of a single hidden-layer complex MLP.

Definition 1. *Affine set of single hidden-layer MLP output. Let A^n be the class of all affine functions from $\mathbb{C}^n$ to $\mathbb{C}$, which is the set of all functions of the form $\sigma(z) = W^T z + \theta$ where W, $z \in \mathbb{C}^n$ and $\theta \in \mathbb{C}$. Here, z corresponds to an n-dimensional complex network input vector, W corresponds to the complex network weight vector from input to the intermediate hidden layer, and θ corresponds to the complex bias. Let Γ be a set of all finite linear combination of affine functions in A^n:*

$$\Gamma = \left\{ f\colon \mathbb{C}^n \to \mathbb{C} : f(z) = \sum_{q=1}^{m} \beta_q \sigma_q(A) = \sum_{q=1}^{m} \beta_q \sigma_q(W_q^T z + \theta_q) \right\} \tag{3.2}$$

Here, m is the number of neurons in the hidden layer, and β_q is the qth complex weight from the hidden layer to the output layer that has a single linear output neuron. The most representative case is when σ_q is the complex squashing function given in definition 3. Thus, Γ is the familiar class of output functions for a single hidden-layer feedforward MLP.

The set of single-hidden-layer MLP output functions Γ in equation 3.2 does not form an algebra, which is required to show that the functions are separating. A set is separable if it contains a countable dense subset (Rudin, 1974). To be an algebra, Γ should be closed under addition, multiplication, and scalar multiplication. To ensure this, an extended set of affine function class is needed to ensure that the set of output functions Γ is closed under multiplication. This extension is possible by considering a class of functions given in equation 3.3:

$$F(z) = \sum_{q=1}^{m} \beta_q \prod_{l=1}^{s_q} \sigma(W_{ql}^T z + \theta_q). \tag{3.3}$$

Here, $\sigma(z)$: $\mathbb{C}^n \to \mathbb{C}$ is the output of a hidden-layer neuron, $F(z)$: $\mathbb{C} \to \mathbb{C}$ is the output of the single output neuron, β_q, $\theta_q \in \mathbb{C}$, W_q, $z \in \mathbb{C}^n$, and m is the number of neurons in the hidden layer. The approximation over the product form given in equation 3.3 naturally yields the single-hidden-layer approximation case as well, which is the special case with $m = 1$. The set of functions in the form of equation 3.3 is denoted as Ψ. Note that $\Gamma \subseteq \Psi$.

In the form given in equation 3.3, note that the complex MLP represents a linear vector mapping by a finite superposition of nonlinear activation functions. The linear vector transform Λ of a vector space V into a vector space V_1 is a mapping of V into V_1 such that $\Lambda(\alpha x + \beta y) = \alpha\Lambda x + \beta\Lambda y$ for all $x, y \in V$ and for all scalars α and β. If V_1 is the field of scalars, as in the case of a complex MLP with a single output neuron, Λ is called a linear functional.

The approximation of continuous linear functionals in $C(I_n)$ is provided by the Riesz representation theorem (Rudin, 1974), while the supporting domain is extended by the Stone-Weierstrass theorem (Rudin, 1991). For the approximation of complex functions with removable singularity that are bounded complex measurable linear functionals in $M(I_n)$, the Riesz representation theorem can be expanded. For complex mappings by superposition of isolated and essential singularities, we use the Laurent series approximation (Silverman, 1975). Generalization of the results to a higher-dimensional range in $\mathbb{C}^s$ and to the multiple hidden-layer MLP structures can be established following the same steps as in Hornik et al. (1989) and are omitted here.

The fundamental feature of a fully complex activation function to enable universal approximation is that it is discriminatory (Cybenko, 1989), which allows the output of a MLP over any compact subset in I_n to be expandable to all of $C(I_n)$ by the Stone-Weierstrass theorem.

Definition 2. *Discriminating function. A function σ is discriminating if for a measure $\mu \in M(I_n)$, $\int_{I_n} \sigma(W^T z + \theta)\, d\mu(z) = 0$ for all $W \in \mathbb{C}^n$ and $\theta \in \mathbb{C}$, implies that $\mu = 0$ (a.e. for a measurable function σ. In other words, σ is* discriminating *if it does not* vanish *anywhere in the domain—in this case, I_n.*

Note that the following theorem on continuous discriminatory functions does not require squashing activation functions in its hidden layer:

Theorem 1. *Let σ: $\mathbb{C}^n \to \mathbb{C}$ be any complex continuous discriminatory function. Then the finite sums of the product of the form $F(z) = \sum_{k=1}^{m} \beta_k \prod_{l=1}^{s_k} \sigma(W_{kl}^T z + \theta_k)$ are dense in $C(I_n)$, that is, $\forall g \in C(I_n)$ and $\varepsilon > 0$, $\exists F(z)$ of the above form such that $|F(z) - g(z)| < \varepsilon$, $\forall z \in I_n$.*

See the proof in the appendix.

The next theorem provides the universal approximation of bounded measurable complex mappings in $M(I_n)$ for those activation functions with removable singularities.

Theorem 2. *Let $\sigma: \mathbb{C}^n \to \mathbb{C}$ be any complex bounded measurable discriminatory function. Then the finite sums of the form of equation 3.3 are dense in $L^1(I_n)$.*

See the proof in the appendix.

For unbounded but measurable tangent function family with isolated singularities, the unbounded singularities pose problems in the Fourier transform representation used in the proof of Theorem 2. However, the isolated and essential singularities can provide uniformly converging approximation in the deleted annulus of the singularity within the radius of convergence defined by the Laurent series. In other words, these ETFs converge with probability 1 within the domain of convergence defined by the shortest radius to the nearest singularity from the center of approximation.

Theorem 3. *Let $\sigma: \mathbb{C}^n \to \mathbb{C}$ be any complex function having isolated and essential singularity. Then the finite sums of the form of equation 3.3 are dense in compact subsets of the analytic deleted neighborhood of the singularity.*

See the proof in the appendix.

Note that the efficiency of MLP using activation functions with essential singularity can be even more powerful than the previous two cases because the deleted neighborhood of essential singularity assumes each complex value infinitely often by Big Picard's theorem (Silverman, 1975). In practice, however, a priori information on the domain of convergence is not necessarily available, and it has been observed that the training and testing results can be more sensitive to parameter initialization choice and to differences in the input noise levels than the fully complex MLP with removable or isolated singularities (Kim, 2002). The characteristics and proper supporting conditions for employing essential singularity require the evaluation of Laurent series approximation that is not always trivial in practical situations as they require intensive numerical computation of large numbers of power series terms.

We have presented three universal approximation theorems using nine continuous, bounded, or unbounded measurable ETFs. Note that with a priori information of the domain and appropriate initialization of random weights to establish a deleted annulus domain, essential singularity can be a powerful activation function as well.

4 Numerical Examples

4.1 Application of Complex MLP. Recent advances in wireless mobile communication and radio navigation have created an ample opportunity for adaptive nonlinear signal processing applications in the complex domain. Arguably, the linearization of power amplifier, in either predistortion or postequalization forms, has been the most frequently encountered application that incorporates adaptive nonlinear signal processing using MLP, recurrent, and radial basis function networks in the complex domain. Examples of their applications in communications are Benvenuto et al. (1991), Benvenuto and Piazza (1992), Chen, Grant, McLaughlin, and Mulgrew (1993), Kechriotis and Manolakos (1994), Ibnkahla and Castanie (1995), You and Hong (1998), Uncini, Vecci, Campolucci, and Piazza (1999), Deng, Sundararajan, and Saratchandran (2000), Langlet, Abdulkader, Roviras, Mallet, and Castanie (2001), Katz (2001), and Park and Jeong (2002). In the applications considered in these references, it is critical to combat the often time-varying nonlinear distortion of signal constellation to match a set of desired target complex vectors. The distortion of the complex constellation is observable in amplitude-to-amplitude and amplitude-to-phase intermodulation-distortion products that hamper the transmission of nonconstant envelope-modulated signals through traveling wave tube or solid-state power amplifiers.

The advantages of a fully complex MLP-based equalizer over the split-complex MLP in robust error performance have been shown, for example, in Kim and Adalı (2000, 2001) and Kim (2002).

Using the simulation of a GPS receiver, it is also shown that a fully complex MLP provides more resilient code and carrier tracking capability than a split-complex MLP when incorporated to predict and cancel the channel fading caused by ionospheric scintillation, known to follow Brownian noise process (Kim, 2002; Kim & Hegarty, 2002).

4.2 Simple Numerical Example. A simple nonlinear system identification example is developed in Kim (2002) to demonstrate the efficiency of fully complex MLP over traditional complex-valued MLP. The input layer of the complex MLPs uses tapped-delay inputs, and the input $x[n]$ is the equally likely binary random sequence $\{1, -1\}$. The system to be identified is given by the following z-transform:

$$H(z) = (1 - 0.3e^{j\pi/6}z^{-1})(1 - 0.3e^{-j\pi/2}z^{-1}).$$

Let the output of $H(z)$ be given by $y_1[n]$; the output $y[n]$ to be identified is obtained through the transformation $y[n] = -0.5x[n] + y_1^3[n]$. The signal constellation for the output $y[n]$ without any noise in the system is shown in Figure 14. Therefore, a deterministic pattern of eight constellation points located both outside and inside the unit circle provides a not-too-trivial system identification example.

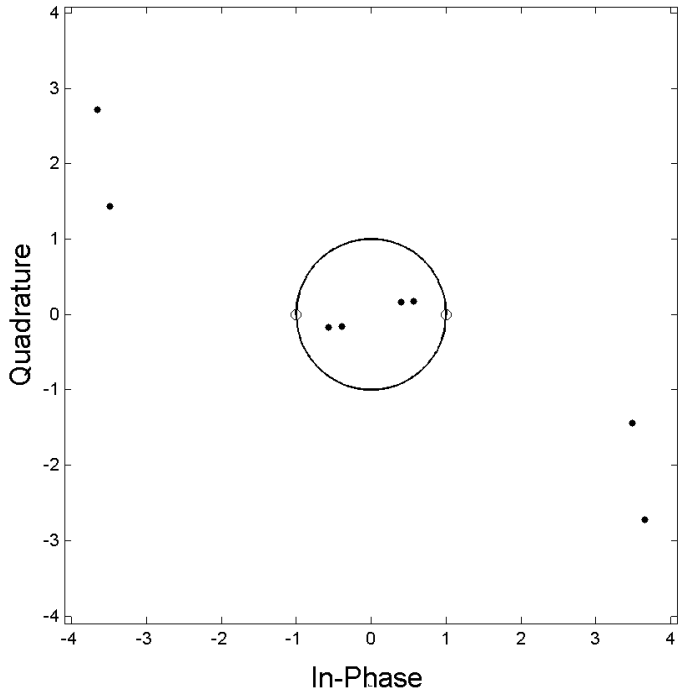

Figure 14: Nonlinear system identification target constellation.

To understand the properties of individual activation functions, the efficiency of standard gradient-descent backpropagation using the best learning parameters is studied. In our simulation example, for comparison, we have defined the best learning rate as the largest possible learning rate without causing a divergence in 50 successive independent trials. Convergence is defined as meeting the preset 1e-9 mean square error (MSE) level within 10,000 training samples. Similar to the learning rate effect, tight initial distribution of complex random weights centered at the origin tends to slow convergence but improves stability, while wide distribution speeds up convergence while also increasing the risk of divergence, except for the case of functions with essential singularity that requires deleted annulus domain. All of these parameters also exhibit dependency to the input and target data distributions and the type of activation function used. Without loss of generality, the best initial radius of random weights, given the best learning rate, is defined as the largest radius that bounds uniformly distributed random complex initial weights centered at the origin, without causing a divergence in 50 successive independent trials.

Table 1: Learning Statistics of Complex-Valued MLPs.

Activation Function	Best Learning Rate	Best Initial Weight Radius	Average	Standard Deviation	Median	Minimum	Maximum
			Convergence in Number of Training Samples (Rounded)				
ARCSINH	0.075	0.1	81	63	70	10	277
ARCSIN	0.1	0.01	87	66	60.5	9	253
ARCTANH	0.4	0.1	148	115	121	14	614
ARCTAN	0.4	0.1	152	146	93	18	647
SIN	0.02	0.02	228	198	169	42	828
SINH	0.04	0.02	235	184	196	29	1121
Split-TANH	0.2	0.2	313	244	254	20	969
GK	0.8	0.1	536	279	543	37	1121
TANH	1.94e-3	0.005	1474	973	1476	138	3995
ARCCOS	0.04	0.01	1521	680	1487	563	3133
TAN	1.09e-4	0.001	2251	2202	1181	203	8912
Hirose	7.03e-4	0.02	3146	1063	2991	1621	6256

To find the best learning rate and initial radius of random weights, an iterative search algorithm is developed. The algorithm starts with a given initial fixed learning rate and an initial radius of weights to test if 50 consecutive successful learning trials for independent random ± 1 input sequences are achievable by meeting a strict 1.0e-9 MSE termination goal. If successful, then the learning rate is doubled to test if new 50 consecutive successful learning trials are still possible and the learning rate is doubled until 50 consecutive successful runs become unsuccessful. When 50 consecutive runs are unsuccessful, then the learning rate is reset to the average between the current one and the previous successful learning rate. Then the test for 50 consecutive successful learning is repeated. There is a lower limit learning rate at 1e-9 to stop and to declare the failure of achieving 50 consecutive successful learning trials at any learning rate. The largest learning rate found during the search is stored with the learning statistics; then the initial radius of weights follows the same binary search procedure to find the best initial radius. The average convergence speed is the final criterion when the set of best learning rate and initial radius pair data is gathered. The MLP structure has three input and three hidden nodes and a single output. The statistics of 50 independent BPSK input data using the best initial learning rate and radius of weights are collected by complex activation function as shown in Table 1.

We can compare the efficiency of each nonlinear activation function in terms of its average convergence characteristics in the fourth column of Table 1. The rows of the table are sorted according to the ascending order of average convergence achieved, such that the top-performing $\operatorname{arcsinh} z$ acti-

vation function is on the top, while the worst-performing Hirose's activation function is placed at the bottom.

The fact that the arcsinh z and arcsin z functions shown in Figures 6 and 7 had the best average convergence characteristics is not surprising. These functions not only provide a consistent circular contour amplitude response while being analytic but also have steadily increasing dynamic ranges. This property tends to avoid the nondiscriminating flat portion of bounded sigmoidal functions, a property that motivated the introduction of a resilient backprop (RPROP) algorithm (Riedmiller & Braun, 1993; Bishop, 1999). Also note that the branch-cut removable singularity did not pose convergence problems in these functions for which we have shown the a.e. universal approximation property in section 3. Following these observations, the average best performance of arctanh z, arctan z, sin z, and sinh z functions followed by arcsinh z and arcsin z functions can be understood by noting the more asymmetrical and squashing type characteristics of these functions.

It may be surprising to see that the split-tanh x and GK's joint-complex functions outperformed tanh z, tan z, and arccos z functions. For tanh z and tan z functions, the periodicity of these functions at every $(1/2 + n)\pi$ on imaginary and real axis, respectively, might have caused the slower convergence. These two functions converged in similar rates with other fully complex functions when the domain shrank to a smaller neighborhood near the unit circle. The slow convergence of arccos z might have been caused by its asymmetric contour as seen in Figure 8, where the singularity is located at 1 on the real axis, not at the origin.

5 Summary

In this article, three types of singularities that elementary transcendental functions possess are first categorized: the removable, isolated, and essential singularities on which three types of approximation theorems are based. For the set of continuous elementary transcendental functions that includes sin z and sinh z, the complex version of the Stone-Weierstrass theorem is applied to show that the class of fully complex MLP output Γ using continuous discriminating activation functions is dense in the space of all continuous complex functions $C(I_n)$ on the n-dimensional complex unit cube I_n. The Riesz representation theorem is used to show that Γ does not vanish at any point in I_n such that every bounded linear functional Λ on $C(I_n)$ has a corresponding finite Borel measure $\mu \in M(I_n)$.

Next, it is noted that those elementary transcendental activation functions with removable and bounded singularities over a bounded domain are complex bounded measurable functions. They are arcsin z, arcsinh z, and arccos z functions for which the extension of Riesz representation and the application of Fourier transform to the discriminating function show that the output functions of MLP using these activation functions are a.e. dense in $C(I_n)$.

The tangent function family including $\tan z$, $\tanh z$, $\arctan z$, and $\operatorname{arctanh} z$ is measurable but not necessarily bounded with isolated singularities. The Laurent series representation that converges uniformly in the deleted neighborhood of the isolated singularity provides the uniform learning capability of MLP employing these functions.

Finally, the essential singularity represented by the base function $e^{1/z}$ is not measurable but has very powerful theoretical representation capability in the deleted neighborhood of singularity suggested by Casorati-Weierstrass and Big Picard's theorems. With a priori knowledge on the working domain of interest where singularities are carefully placed outside the domain, the essential singularity functions are capable of uniform approximation. In practice, however, the strategic placement of singularity is not obvious, and the dense representation of any complex number infinitely often in the neighborhood of essential singularity, as stated in Big Picard's theorem, tends to show high sensitivity in the initial random weight distribution and the size of learning rate. The study of essential singularity remains a good subject for further research.

To summarize, three theoretical existence proofs are established for the approximation capability of the fully complex MLP. First, it is shown that a fully complex MLP with a continuous nonlinear activation function is capable of universal approximation of any continuous mapping over a compact set in $\mathbb{C}^n$. Second, we prove that a fully complex MLP with a bounded measurable nonlinear activation function can provide universal approximation of any measurable mapping a.e. over a compact set in $\mathbb{C}^n$. The third main result shows that a fully complex MLP with an unbounded measurable nonlinear activation function with isolated singularity or a nonmeasurable nonlinear activation function with essential singularity can provide uniform approximation of any nonlinear mapping a.e. over a deleted annulus of singularity. If multiple singularities are included, then the radius of convergence is the shortest distance to a singularity from the origin.

In this article, we have studied the approximation capability of the fully complex MLP. Further areas for research include the challenging and rich problems of improving the training of these fully complex MLPs, by both first-order derivatives as in backpropagation and using higher-order derivatives (Haykin, 1999). These problems include the application of real, imaginary, and complex learning rates, use of adaptive learning rates, and direction-search-type algorithms in the complex domain, as discussed in Kim (2002), as well as further investigation of the properties of different activation functions, such as those with essential singularities.

Appendix

Definition 3. *Complex squashing function. A complex measurable function* Ψ: $H \to D$, *where* H *and* D *are bounded subsets of* $\mathbb{C}$, *is a squashing function if its*

magnitude is nondecreasing, $\lim_{\mathrm{Inf}|z|, z\in H} |\Psi(z)| = c$, *for* $0 \le d < c < \infty$ *and* $\lim_{|z|\to 0} |\Psi(z)| = d$.

Definition 4. *Positive and complex measures (Rudin, 1974):*

- *A positive measure is a function* μ, *defined on a* σ*-algebra* $\Re$, *whose range is in* $[0, \infty]$ *and is countably additive. This means that if* $\{A_i\}$ *is a disjoint countable collection of members of* $\Re$, *then*

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$

- *A* measure space *is a measurable space that has a positive measure defined on the* σ*-algebra of its measurable sets.*
- *A* complex measure *is a complex-valued countably additive function defined on a* σ*-algebra.*

Definition 5. ρ*-dense metric space (Hornik* et al.*, 1989). A subset S of a metric space* (X, ρ) *is* ρ-dense *in a set T if for every* $\varepsilon > 0$ *and for every* $t \in T$, *there is an* $s \in S$ *such that* $\rho(s, t) < \varepsilon$.

Definition 6. *Uniform dense space (Hornik et al., 1989). A subset S of* $C(I_n)$ *is said to be uniformly dense on compacta in* $C(I_n)$ *if for every compact subset* $K \subset \mathbb{C}^n$, *S is* ρ*-dense in* $C(I_n)$, *where for* $f, g \in C(I_n)$, $\rho(f, g) \equiv \sup_{z\in K} |f(z) - g(z)|$. *A sequence of functions* $\{f_n\}$ *converges to a function f uniformly on compacta if for all compact* $K \subset \mathbb{C}^n$, $\rho(f_n, f) \to 0$ *as* $n \to \infty$.

Proof of Theorem 1. First, we show the denseness of $F(z)$ in $C(I_n)$ over a compact subset K of I_n. Then we apply the Stone-Weierstrass theorem that requires the following four conditions for the uniform closure of subspace $\Psi \in C(I_n)$ to be equivalent to $C(I_n)$ over K. Consequently, the Stone-Weierstrass theorem says that Ψ is ρ-dense in $C(I_n)$ on K (Ash, 1972):

1. Ψ is a closed subalgebra of $C(I_n)$.
2. Ψ is self-adjoint (i.e., $\bar{f} \in \Psi$ for all $f \in \Psi$).
3. Ψ separates points on $C(I_n)$.
4. At every $p \in I_n$, $f(p) \neq 0$ for some $f \in \Psi$.

Step 1: Let $K \subset I_n$ be any compact set. For any continuous σ over K, $F(z) \in \Psi$ is obviously a closed subalgebra Ψ of $C(I_n)$ over K, that is, Ψ is closed under addition, multiplication, and scalar multiplication that includes the single hidden-layer output in equation 3.2 as a special case. Since real-valued continuous functions are in Ψ, and Ψ is an algebra over symmetric compact set K, it is trivial that Ψ is self-adjoint, since if $f \in \Psi$, then $\bar{f} \in \Psi$ by

replacing the complex weights to represent f by their conjugate weights. Ψ also separates since if $x, y \in K$, $x \neq y$, then there should be a $\{W, \theta\}$ such that $\sigma(W^T x + \theta) \neq \sigma(W^T y + \theta)$. To see this, pick $u, v \in \mathbb{C}$, $u \neq v$ such that $F(u) \neq F(v)$. Now pick $\{\hat{W}, \hat{\theta}\}$ to satisfy $\rho(\hat{W}^T x + \hat{\theta}) = u$ and $\rho(\hat{W}^T y + \hat{\theta}) = v$. Then $F(\rho(\hat{W}^T x + \hat{\theta})) \neq F(\rho(\hat{W}^T y + \hat{\theta}))$. Therefore, Ψ is separating on K. To satisfy condition 4, we use the Riesz representation theorem where every bounded linear functional Λ on $C(I_n)$, has a corresponding finite Borel measure $\mu \in M(I_n)$ such that $\Lambda f = \int_{I_n} f \, d\mu$, for all $f \in C(I_n)$. Since $\sigma(W^T z + \theta)$ is discriminating in Ψ for $\forall W \in K$, $F(z)$ cannot vanish at any point in K unless the measure μ itself is a zero measure. Thus, condition 4 is satisfied.

Step 2: The four required conditions for the Stone-Weierstrass theorem are now satisfied. This implies that Ψ is dense in the space of complex continuous functions over a compact set K, but since K is arbitrary, the single-hidden complex MLP output functions $G(\in \Gamma \subset \Psi \subset C(I_n))$ must be dense in $C(I_n)$. In other words, Ψ is ρ-dense in $C(I_n)$, as in definition 5. It is also known that the supremum norm in $L^1(\mu)$ can be generalized to $L^p(\mu)$-norm with $0 < p < \infty$ (see theorem 4.3.14 of Ash, 1972).

Proof of Theorem 2. The Riesz representation can be expanded into bounded linear functionals. Let μ be a complex measure on a σ-algebra in X. Then there is a complex measurable Borel function h with $|h| = 1$ such that $d\mu = h d|\mu|$. This enables new integration with respect to a complex measure μ by the formula $\int f \, d\mu = \int fh \, d|\mu|$. Therefore, to show that $\int_K \sigma(W^T z + \theta) \, d\mu(z)$ is discriminatory is now equivalent to showing that $\int_K \sigma(W^T z + \theta) h(z) |\mu(z)|$, $\forall W$, θ implies that $h(\cdot) = 0$ a.e. As shown by Cybenko, this is equivalent to obtaining the Fourier transform in the complex domain $F(\exp(imt)) = \int_K \exp(imt) h(t) \, dt = 0$ for all m to yield that the Fourier transform of h is 0. Therefore h itself is 0, showing that F does not vanish on K, as shown in theorem 1.

Proof of Theorem 3. The elementary transcendental functions having isolated and essential singularities are analytic in the deleted neighborhood of the singularities. For a single variable function f having an isolated singularity at $z = z_0$, $f(z)$ has a pole at $z = z_0$. At the pole, $f(z)$ may be expressed as a power series $f(z) = \sum_{n=-k}^{\infty} b_n (z - z_0)^n$, where k is the order of the pole that is valid in a deleted neighborhood of z_0. More generally, using Laurent's theorem (Silverman, 1975) if $f(z)$ is analytic in the annulus $R_1 < |z - z_0| < R_2$, where R_2 is the minimum distance among singularities to the origin, then the Laurent series pole representation $f(z) = \sum_{n=-\infty}^{\infty} a_n (z - z_0)^n$ is valid, that is, converges uniformly, throughout the annulus and the coefficients are given by $a_n = \frac{1}{2\pi i} \int_C \frac{f(\zeta)}{(\zeta - z_0)^{n+1}} \, d\zeta$, where C is any simple closed contour contained in the annulus that makes a complete counterclockwise revolution about the point z_0. For essential singularity, the uniform convergence of power series approximation of any complex number is summarized by the

Casorati-Weierstrass theorem (Silverman, 1975) where $f(z)$ with isolated singularities comes arbitrarily close to every complex value in each deleted neighborhood of z_0. The finite linear sum of isolated and essential singularity in the feedforward complex MLP therefore can represent any complex value arbitrary closely in the annulus $R_1 < |z - z_0| < R_2$.

Acknowledgments

This work was supported in part by the National Science Foundation Career Award, NSF NCR-9703161.

References

Ash, B. R. (1972). *Real analysis and probability*. New York: Academic Press.

Benvenuto, N., Marchesi, M., Piazza, F., & Uncini, A. (1991). Non linear satellite radio links equalized using blind neural networks. In *Proc. of International Conference on Acoustics Speech and Signal Processing* (Vol. 3, pp. 1521–1524). Piscataway, NJ: IEEE Press.

Benvenuto, N., & Piazza, F. (1992). On the complex backpropagation algorithm. *IEEE Trans. on Signal Processing, 40*, 967–969.

Bishop, C. (1999). *Neural networks for pattern recognition*. New York: Oxford University Press.

Chen, S., Grant, P. M., McLaughlin, S., & Mulgrew, B. (1993). Complex-valued radial basis function networks. In *Proc. of Third IEEE International Conference on Artificial Neural Networks* (pp. 148–152). Piscataway, NJ: IEEE Press.

Churchill, R., & Brown, J. (1992). *Complex variables and applications* (5th ed.). Seoul, Korea: McGraw-Hill.

Clarke, T. (1990). Generalization of neural network to the complex plane. In *Proc. of International Joint Conference on Neural Networks* (Vol. 2, pp. 435–440).

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

Deng, J., Sundararajan, N., & Saratchandran, P. (2000). Communication channel equalization using complex-valued minimal radial basis functions neural network. In *Proc. of IEEE IJCNN 2000* (Vol. 5, pp. 372–377). Piscataway, NJ: IEEE Press.

Funahashi, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks, 2*, 183–192.

Georgiou, G., & Koutsougeras, C. (1992). Complex backpropagation. *IEEE Trans. on Circuits and Systems II, 39*, 330–334.

Haykin, S. (1999). *Neural networks: A comprehensive foundation* (2nd ed.). New York: Macmillan.

Hirose, A. (1992). Continuous complex-valued back-propagation learning. *Electronics Letters, 28*, 1854–1855.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*, 359–366.

Ibnkahla, M., & Castanie, F. (1995). Vector neural networks for digital satellite communications. In *Proc. of International Conference on Communications* (Vol. 3, pp. 1865–1869). Piscataway, NJ: IEEE Press.

Katz, A. (2001, December). Linearization: Reducing distortion in power amplifiers. *IEEE Microwave Magazine*, December, 37–49.

Kechriotis, G., & Manolakos, E. (1994). Training fully recurrent neural networks with complex weights. *IEEE Trans. on Circuits and Systems—II: Analog and Digital Signal Processing, 41*, 235–238.

Kim, T. (2002). *Fully complex multilayer perceptron and its application to communications*. Unpublished doctoral dissertation, University of Maryland.

Kim, T., & Adalı, T. (2000). Fully complex backpropagation for constant envelop signal processing. In *Proc. of IEEE Workshop on Neural Networks for Signal Processing* (NNSP) (pp. 231–240). Piscataway, NJ: IEEE Press.

Kim, T., & Adalı, T. (2001). Complex backpropagation neural network using elementary transcendental activation functions. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP) (Vol. 2). Piscataway, NJ: IEEE Press.

Kim, T., & Adalı, T. (2002). Fully-complex multilayer perceptron for nonlinear signal processing. *Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology, 32*, 29–43.

Kim, T., & Hegarty, C. (2002). Simulated nonlinear prediction and cancellation of ionospheric scintillation in GPS/WAAS software receivers. In *Ionospheric Effects Symposium 2002(IES2002)* (pp. 5B9-1–5B9-12). Alexandria, VA: JMG Associates.

Langlet, F., Abdulkader, H., Roviras, D., Mallet, A., & Castanie, F. (2001). Adaptive predistortion for solid state power amplifier using multi-layer perceptron. In *IEEE Globecom'01* (Vol. 1, pp. 325–329). Piscataway, NJ: IEEE Press.

Leung, H., & Haykin, S. (1991). The complex backpropagation algorithm. *IEEE Trans. on Signal Proc., 3*, 2101–2104.

Mandic, D., & Chambers, J. (2001). *Recurrent neural networks for prediction*. New York: Wiley.

Needham, T. (2000). *Visual complex analysis*. New York: Oxford University Press.

Park, D-C., & Jeong, T-K. (2002). Complex-bilinear recurrent neural network for equalization of a digital satellite channel. *IEEE Trans. on Neural Networks, 13*, 722–725.

Riedmiller, M., & Braun, H. (1993). A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In *Proc. of IEEE International Conference on Neural Network, 1*, 586–591. Piscataway, NJ: IEEE Press.

Rudin, W. (1974). *Real and complex analysis* (2nd ed.). New York: McGraw-Hill.

Rudin, W. (1991). *Functional analysis* (2nd ed.). New York: McGraw-Hill.

Silverman, H. (1975). *Complex variables*. Boston: Houghton Mifflin.

Uncini, A., Vecci, L., Campolucci, P., & Piazza, F. (1999). Complex-valued neural networks with adaptive spline activation functions. *IEEE Trans. on Signal Processing, 47*, 505–514.

You, C., & Hong, D. (1998). Nonlinear blind equalization schemes using complex-valued multilayer feedforward neural networks. *IEEE Trans. on Neural Networks, 9*, 1442–1455.

Received June 26, 2002; accepted January 6, 2003.