Approximation by Fully Complex MLP Using Elementary Transcendental Activation Functions

Taehwan Kim[†,] and Tülay Adali[*]*

[†]Center for Advanced Aviation System Development
The MITRE Corporation
McLean, Virginia 22102 U.S.A.

[*]Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County
Baltimore, Maryland 21250 U.S.A.

ABSTRACT

Recently, we have presented 'fully' complex multi-layer perceptrons (MLPs) using a subset of complex elementary transcendental functions as the nonlinear activation functions. These functions jointly process the in-phase (I) and quadrature (Q) components of data while taking full advantage of well-defined gradients in the error back-propagation. In this paper, the characteristics of these elementary transcendental functions are categorized and their common *almost everywhere* (*a.e.*) bounded and analytic properties are investigated. More importantly, it is proved that fully complex MLPs are *a.e.* convergent and therefore are capable of universally approximating any nonlinear complex mapping to an arbitrary accuracy. Numerical examples demonstrate the benefit of *isolated essential singularity* included in a subgroup of elementary transcendental functions in achieving arbitrarily close approximation to the desired mapping.

1. INTRODUCTION

One of the early efforts to process complex data using neural networks (NNs) can be found in Clarke's 1990 paper [1] that generalized the real-valued activation function (AF) into the complex-valued one. He extended tanh x, $x\in\mathbf{R}$ into tanh $z = (e^z-e^{-z})/(e^z+e^{-z})$, $z\in\mathbf{C}$ where he noted that the extension destroys the boundedness property of the real-valued sigmoid function. He pointed out that it is impossible to avoid this problem as a consequence of Liouville's theorem which states that the only bounded differentiable (analytic) functions defined for the entire complex domain **C** are constant functions [2]. Clarke noted that the singularities that occur in a limited region of the complex plane for tanh z, e.g., at $z = (1/2+n)\pi i$, where n is an integer and $i=\sqrt{-1}$, are not necessarily a problem and indeed lead to interesting neural behavior (see Figure 3). He didn't present the derivation of the complex back-propagation algorithm but noted that to keep the (activation) function analytic, one cannot avoid unbounded functions.

Clarke's observation on the inevitable unbounded nature of analytic functions in **C** was compromised in 1991 by Leung and Haykin [3] and Benvenuto *et al.* [4] who almost simultaneously introduced "ad-hoc" extension of the real-valued back-propagation (BP) algorithm to complex signals. Both approaches employed a pair of tanh x, $x \in \mathbf{R}$ (hence the name '*split*' complex AF) to marginally process the I and Q components of weighted sum of input data as shown in Eqn. (1):

$$f(z_k) = f_R(\mathrm{Re}(z_k)) + i f_I(\mathrm{Im}(z_k)), \;\; z_k = \sum_l W_{kl} X_l \tag{1}$$

where z_k is the input to the k-th neuron obtained from the complex weighted sum of the hidden-layer complex weights W_k and the input data X. Here, the real and imaginary components of z_k, i.e., $\mathrm{Re}(z_k)$ and $\mathrm{Im}(z_k)$, are split and fed through the real-valued $f_R(x) = f_I(x) = \tanh(x)$, $x \in \mathbf{R}$ (see Figure 10). Even though bounded, the complex activation function defined in this way is not analytic and the back-propagation from the output layer down to hidden layer also takes split paths through disjoint real-valued gradients. As would be expected, such a scheme will not be efficient when learning nonlinear mappings of complex input/output pairs.

In 1992, Georgiou and Koutsougeras [5] and Hirose [6] proposed joint nonlinear complex AFs that process the I and Q components as shown in Eqns. (2) and (3), respectively.

$$f(z) = z / (c + |z| / r) \tag{2}$$

$$f(s \exp[i\beta]) = \tanh(s/m) \exp[i\beta] \tag{3}$$

where, c, r, and m are real positive constants and $z = s\,exp[i\beta]$. However, these functions preserve the phase, thus are difficult to learn the phase variations between the input and the target in pattern classification tasks where the input layer does not include time delays. Also, as shown in our previous paper in [7], their performance is usually not as good as the fully complex MLPs employing elementary transcendental functions (ETFs). Note that even though they are bounded everywhere in **C**, neither of them is analytic therefore does not qualify as a fully complex activation function.

We have recently identified and implemented nine ETFs as fully complex nonlinear activation functions for complex MLPs [7]. They benefit from a compact form of fully complex back-propagation satisfied by *almost everywhere* (*a.e.*) bounded and analytic property. In this paper, the ETFs are categorized into two groups to establish the proof that the fully complex MLPs are capable of universally approximating any complex mapping with arbitrary accuracy, and they converge *a.e.* in a bounded domain of interest.

2. ELEMENTARY TRANSCENDENTAL FUNCTIONS

We have implemented the following ETFs in nonlinear satellite channel equalization examples. Figures 1 through 9 show the magnitude of these ETFs in the vicinity of unit circle. Figure 10 shows the bounded magnitude of split complex tanh *x* AF for a comparison. These figures show two distinctive patterns of inevitable unboundedness property when real-valued ETFs are generalized into complex-valued ETFs.

- <u>Circular functions</u>

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

- <u>Inverse circular functions</u>

  $\arctan z = \int_0^z \dfrac{dt}{1+t^2}, \left\{ \dfrac{d}{dz} \arctan z = \dfrac{1}{1+z^2} \right\}$, $\arcsin z = \int_0^z \dfrac{dt}{(1-t)^{1/2}}, \left\{ \dfrac{d}{dz} \arcsin z = (1-z^2)^{-1/2} \right\}$,

  $\arccos z = \int_z^1 \dfrac{dt}{(1-t^2)^{1/2}}, \left\{ \dfrac{d}{dz} \arccos z = -(1-z^2)^{-1/2} \right\}$

- <u>Hyperbolic functions</u>

  $\tanh z = \dfrac{\sinh z}{\cosh z} = \dfrac{e^z - e^{-z}}{e^z + e^{-z}}, \left\{ \dfrac{d}{dz} \tanh z = \sec h^2 z \right\}$, $\sinh z = \dfrac{e^z - e^{-z}}{2}, \left\{ \dfrac{d}{dz} \sinh z = \cosh z \right\}$

- <u>Inverse hyperbolic functions</u>

  $\operatorname{arc\,tanh} z = \int_0^z \dfrac{dt}{1-t^2}, \left\{ \dfrac{d}{dz} \operatorname{arc\,tanh} z = (1-z^2)^{-1} \right\}$,

  $\operatorname{arcsinh} z = \int_0^z \dfrac{dt}{(1+t^2)^{1/2}}, \left\{ \dfrac{d}{dz} \operatorname{a\,sinh} z = (1+z^2)^{-1} \right\}$

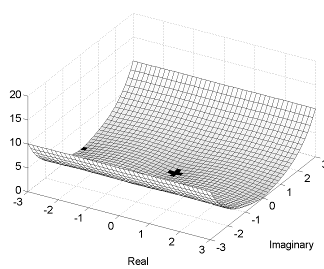

Figure 1. Magnitude of sin *z*

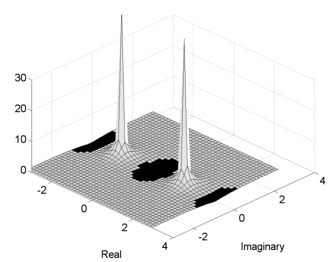

Figure 2. Magnitude of tan *z*

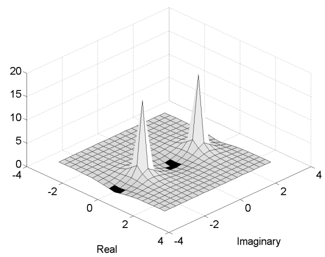

Figure 3. Magnitude of tanh *z*

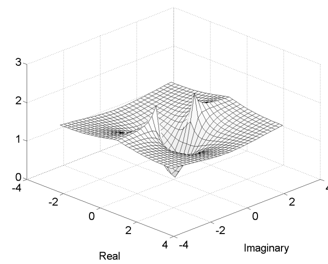

Figure 4. Magnitude of arctan *z*

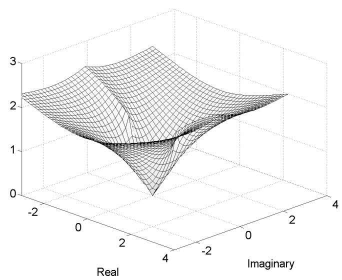

Figure 5. Magnitude of arcsin *z*

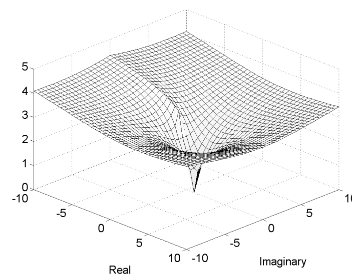

Figure 6. Magnitude of arccos *z*

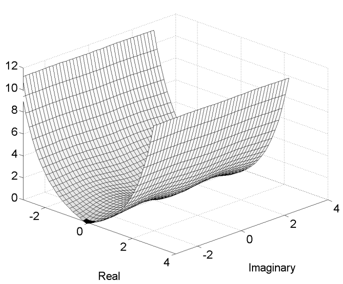

Figure 7. Magnitude of sinh *z*

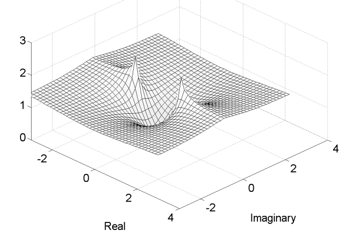

Figure 8. Magnitude of arctanh *z*

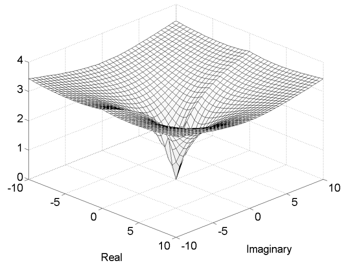

Figure 9. Magnitude of arcsinh *z*

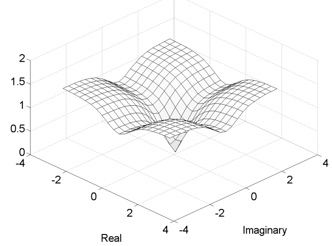

Figure 10. Magnitude of split tanh *x*

First category of unboundedness is observable in the tangent function family (tan z, tanh z, arctan z, and arctanh z) that possesses finite number of point set discontinuities in a bounded domain. A single-valued function is said to have a *singularity* at a point if the function is not analytic at the point. If the function is analytic in some deleted neighborhood of the point, then the singularity is said to be *isolated*. If $f(z)$ has an isolated singularity at z_0, the singularity is said to be *removable* if $\lim_{z\to z_0} f(z)$ exists. However, if $\lim_{z\to z_0} f(z) \to \infty$, while $f(z)$ is analytic in a deleted neighborhood of $z = z_0$ (e.g., $f(z) = 1/z$), then $f(z)$ is not removable but has a pole at $z = z_0$. At the pole, the function can be expressed as $f(z) = \sum_{n=-k}^{\infty} b_n (z - z_0)^n$, where k is the order of the pole (see p. 236 [2]). Note that arctan z has isolated singularity at $\pm i$ and arctanh z has isolated singularity at ± 1. Furthermore, an isolated singularity that is neither removable nor a pole is said to be an *isolated essential singularity*. Note that tanh z has isolated essential singularities at every $(1/2+n)\pi i$, $n \in \mathbf{N}$, since it is asymptotically $+\infty$ as $(1/2+n)\pi i$ is approached from below and to $-\infty$ as $(1/2+n)\pi i$ is approached from above along the imaginary axis as shown in Figures 11 and 12. Similarly, tan z has isolated essential singularities at every $(1/2+n)\pi$.

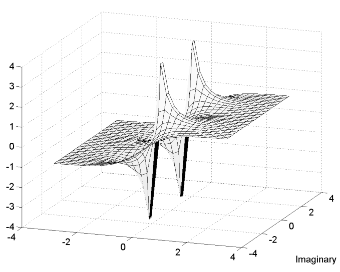

Figure 11. Real part of tanh z

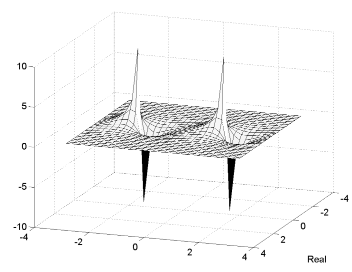

Figure 12. Imaginary part of tanh z

Second category of unboundedness is observable in the inverse sine and cosine family including arcsin z, arccos z, and arcsinh z functions that exhibit unbounded but decreasing rate of magnitude growth as they move away from the origin, as shown in Figures 5, 6, and 9. The inverse sine and cosine functions also include uncontinuous *branch* points outside of the unit circle along either real or imaginary axis. This branch type discontinuity results from the definition of the integral where the branch should not be crossed. The sine function family including sin z and sinh z grow unboundedly with increasing rate of magnitude growth in parallel to either real or imaginary axis, as shown in Figures 1 and 7, respectively.

Therefore, for the tangent function family, it is evident that they are continuous and analytic *a.e.* Note that these functions do not meet the usual definition of a squashing function where the function output should be bounded. For the rest of the ETFs, it is also clear that once the domain of interest is bounded, which represents almost all practical engineering applications using finite amplitude signals, they are fully bounded within the domain. These functions meet the definition of a squashing function even though they are not defined and therefore not analytic at the branch points. However, they still meet the *a.e.* bounded and analytic conditions since the branch is a set of measure zero in the complex plane **C**.

3. UNIVERSAL APPROXIMATION OF FULLY COMPLEX MLP

The squashing function characteristic of inverse sine, cosine, and regular and hyper sine functions is necessary in proving the universal approximation capability of fully complex MLP using Mergelyan's theorem [9]. The same conclusion can be obtained for the tan *z* and tanh *z* AFs with isolated essential singuralities using Casorati-Weierstrass theorem. For arctan *z* and arctanh *z* AFs having isolated singularities, Laurrent's theorem is applicable for a similar but a weaker form of the universal approximation.

First set of theorems are derived for the non-tangent squashing funtion ETFs where we closely follow the proof used in the universal approximation theorem by Hornik *et al.*[8] Several basic definitions provided in [8] are generalized to **C**. Note that a complex measure is a complex-valued countably additive function defined on a σ-algebra [9].

Definition 1 (Complex Squashing Function): A complex-valued measurable function Ψ: $H \rightarrow D$, where H and D are bounded subsets of **C**, is a squashing function if its magnitude is non-decreasing, $\lim_{\mathrm{Inf}|z|, z\in H} |\Psi(z)| = c, 0 < c < \infty$, and $\lim_{|z|\to 0}|\Psi(z)| = 0$.

Definition 2 (Affine Functions): For any $r\in$ **N**={1,2, …}, $\mathbf{A}^r$ is the set of all affine functions from $\mathbf{C}^r$ to **C**, i.e., the set of all functions of the form **A**(***z***) = $\boldsymbol{W}\bullet\mathbf{z} + b$ where $\boldsymbol{W}$ and $\boldsymbol{z}$ are vectors in $\mathbf{C}^r$, and "•" denotes the dot product of complex vectors, and *b* is a complex scalar.

Definition 3. For any complex measurable function $G(\cdot)$ mapping **C** to **C** and $r\in$ **N**, let $\Sigma^r(G)$ be the class of functions

$$\{f : \mathbf{C}^r\to\mathbf{C} : \ f(z) = \sum_{j=1}^{q} \beta_j G(A_j(z)),\ z\in \mathbf{C}^r,\ \beta_j \in \mathbf{C},\ A_j \in \mathbf{A}^r,\ q = 1,2,\cdots\}.$$

Definition 4. For any complex measurable function $G(\cdot)$ mapping **C** to **C** and $r\in$ **N**, let $\Sigma\Pi^r(G)$ be the class of functions

$$\{f:\mathbf{C}^r\to\mathbf{C}:\ f(z) = \sum_{j=1}^{q} \beta_j \prod_{k}^{l_k} G(A_{jk}(z)),\ z\in \mathbf{C}^r,\ \beta_j \in \mathbf{C},\ A_{jk} \in \mathbf{A}^r,\ l_j, q = 1,2,\cdots\}.$$

Definition 5. Let $C(I)^r$ be the set of continuous functions from $\mathbf{C}^r$ to $\mathbf{C}$, and let M^r be the set of all Borel measurable functions from $\mathbf{C}^r$ to $\mathbf{C}$. We denote the Borel σ-field of $\mathbf{C}^r$ as B^r.

The classes $\Sigma^r(G)$ and $\Sigma\Pi^r(G)$ belong to M^r for any Borel measurable G. When G is continuous, $\Sigma^r(G)$ and $\Sigma\Pi^r(G)$ belong to $C(I)^r$. The class $C(I)^r$ is a subset of M^r. For approximation between two complex functions f and g belonging to $C(I)^r$ or M^r, 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.

Definition 6. A subset S of $C(I)^r$ is said to be uniformly dense on compacta in $C(I)^r$ if for every compact subset $K \subset \mathbf{C}^r$, S is ρ_K – dense in $C(I)^r$, where for f, g $\in$ $C(I)^r$ $\rho_K(f,g) \equiv \mathrm{Sup}_{z\in K}\left|f(z)-g(z)\right|$. A sequence of functions $\{f_n\}$ converges to a function f uniformly on compacta if for all compact $K \subset \mathbf{C}^r$, $\rho_K(f_n, f) \to 0$ as $n \to \infty$.

The first main result relying on the uniform convergence on compacta through ρ_K – denseness in $C(I)^r$ follows. Here, it is important to assume that any compact set K does not separate the plane $\mathbf{C}$ (i.e., the complement of K is connected), which is guaranteed by the embedded sequence $\{K_n\}$ of compact sets as shown in Theorem 13.3 of Rudin [9]. Also note that $H(\Omega)$ indicates the class of all holomorphic (or analytic) functions for every $z_0 \in \Omega$.

Theorem 1. Let G be any complex continuous function on a compact set S which is analytic in the interior of S. Then $\Sigma\Pi^r(G)$ is uniformly dense on compacta in $C(I)^r$.

Proof. Let $K \subset \mathbf{C}^r$ be a compact set. For any G, $\Sigma\Pi^r(G)$ is obviously a complex algebra on K, i.e., it is closed under associative and distributive multiplication as well as scalar multiplication. It is also separating, since if z_1, $z_2 \in K$, $z_1 \neq z_2$ then there is an $A \in \mathbf{A^r}$, such that $G(A(z_1)) \neq G(A(z_2))$. By Mergelyan's theorem (Theorem 20.5, p. 423, [9]), it is known that there exists a polynomial P such that $|P(z) - \Sigma\Pi^r(G)| < \varepsilon$ for all $z \in K$ and for any given $\varepsilon > 0$.

In other words, $\Sigma\Pi$ feed-forward networks are capable of approximating any complex continuous function over a compact set to an arbitrary accuracy. The continuous function condition can be relaxed to show that the *a.e.* analytic condition is sufficient using a probability measure $\mu(\mathbf{C}^r, B^r)$.

Definition 7. Let μ be a probability measure on $(\mathbf{C}^r, B^r)$. If f and g belong to M^r, we say they are μ–equivalent if $\mu\{z \in \mathbf{C}^r: f(z) = g(z)\} = 1$.

Definition 8. Given a probability measure μ on $(\mathbf{C}^r, B^r)$, define the metric ρ_μ from $M^r \times M^r$ to $\mathbf{R}^+$ by $\rho_\mu(f,g) = \inf\{\varepsilon > 0 : \mu\{z : \left|f(z) - g(z)\right| > \varepsilon\} < \varepsilon\}$.

Note that two functions are close in this metric if and only if there is a small probability that they differ significantly. The following lemma [8] relates the uniform convergence of functions to ρ_μ −convergence

Lemma 1. If $\{f_n\}$ is a sequence of functions in M^r that converges uniformly on compacta to the function f then $\rho_\mu(f_n, f) \to 0$.

Theorem 2. For every continuous nonconstant function G, every r, and every probability measure μ on $(\mathbf{C}^r, B^r)$, $\Sigma\Pi^r(G)$ is ρ_μ −dense in M^r.

Theorem 3. (Universal approximation of fully complex MLP) For every squashing function Ψ, every r, and every probability measure μ on $(\mathbf{C}^r, B^r)$, $\Sigma\Pi^r(G)$, and its simpler case $\Sigma^r(G)$, are uniformly dense on compacta in $C(I)^r$ and ρ_μ −dense in M^r.

The construction above that leads to Theorem 3 follows the same steps as in [8] using a complex measure μ instead. It is also possible to arrive at the same conclusion through use of trigonometric polynomial approximation based on the Riesz-Fischer theorem (Chapter 4, [9]) instead of the real-valued cosine squashing function in the proof of Lemma A.3 [8] that leads to the proof of Theorem 3.

On the other hand, the MLPs using tan z and tanh z functions converge arbitrary close to any complex function in the deleted neighborhood of singularity. This is because they have isolated essential singularities for which Casorati-Weierstrass theorem (*Theorem 9.3*, p.237 [2]) is applicable.

Theorem (*Casorati-Weierstrass*). *If* $f(z)$ *has an isolated essential singularity at* $z=z_0$ *then* $f(z)$ *comes arbitrarily close to every complex value in each deleted neighborhood of* z_0.

More general result is obtained in big Picard theorem [9] that has a powerful statement that, with one possible exception, $f(z)$ assumes each complex value infinitely many times in the deleted neighborhood.

Since isolated essential singularity provides the arbitrarily closeness to every complex value in the deleted neighborhood, the Laurent series expansion $f(z) = \sum_{n=-\infty}^{\infty} a_n (z - z_0)^n$ in the annulus of singularity is valid and it converges uniformly on a compact set. Therefore, a corollary of Theorem 2 can be established for the *a.e.* continuous and analytic tan z and tanh z MLPs.

Corollary 1. For complex functions G having isolated essential singularity, every r, and every probability measure μ on $(\mathbf{C}^r, B^r)$, $\Sigma\Pi^r(G)$ is ρ_μ −dense *almost everywhere* in M^r.

Finally, for arctan z and arctanh z functions having isolated singularities (but not essential), we can use the fact that they have poles at the singularities (see Theorem 9.2 [2]). Then they may be expressed in the truncated form of

Laurent series $f(z)=\sum_{n=-k}^{\infty} b_n (z-z_0)^n$, where k is the order of the pole. Even though Laurent series converges uniformly on a compact set, the pole representation does not satisfy the uniform convergence. Therefore, universal approximation theorem cannot be applied to inverse tangent functions, even though the practical application of these functions can still provide adequate performance in many practical applications requiring only finite precision.

4. NUMERICAL EXAMPLES

A relatively simple numerical example can demonstrate the utility of isolated essential singularity suggested by the big Picard theorem where tan z and tanh z can assume any arbitrary complex value infinitely often in the deleted neighborhood of singularity. A mild third order nonlinear channel is modeled to distort the quadrature phase shift keying (QPSK) signal generated from a random uniform sequence $\{\pm 1, \pm i\}$. Eqn. 4 shows the mild nonlinear channel distortion where the first linear component provides 45° constellation rotation while the third component provides a small amount of nonlinear rotation on each constellation point, as shown in Figure 14.

$$\begin{aligned} y(n) &= (0.7071+0.7071i)x(n)+(0.1-0.1i)x(n-1) \\ &\quad +0.2\{(0.7071+0.7071i)x(n)-(0.1-0.1i)x(n-1)\}^3 \end{aligned} \tag{4}$$

Figures 15 and 16 show the restored QPSK constellation by a fully complex MLP using tanh z and an MLP using split tanh x activation functions, both with similar complexity. The training curves given in Figures 17 and 18 along with the restored constellations demonstrate the advantage of the tanh z nonlinearity, i.e., of the isolated essential singularity's 'arbitrarily close and infinitely often' approximation characteristics. MLPs using other fully complex activation functions did not demonstrate the efficient learning behavior of tanh z for this mapping task either. All structures tested had similar complexity.

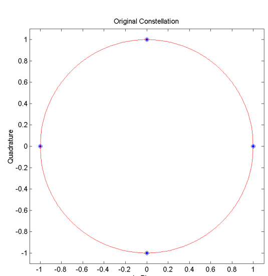

Figure 13. Original QPSK Constellation

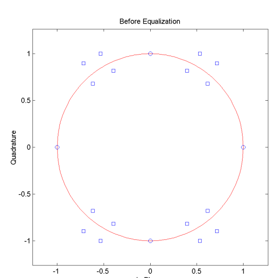

Figure 14. Distorted QPSK Constellation

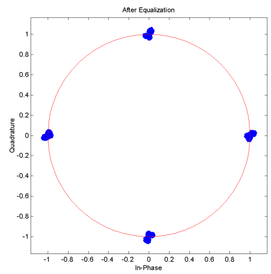

Figure 15. Restored by tanh *z* MLP

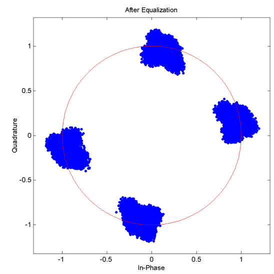

Figure 16. Restored by split tanh *x* MLP

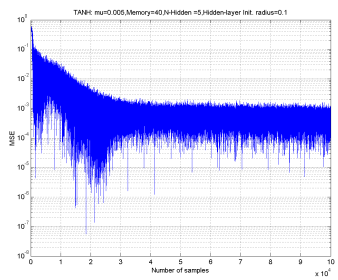

Figure 17. Training MSE of tanh *z* MLP

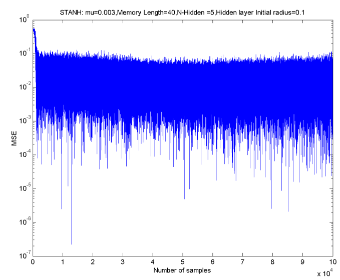

Figure 18. Training MSE of split tanh MLP

REFERENCES

[1] Clarke T., "Generalization of Neural Network to the Complex Plane", *Proc. of IJCNN*, Vol.2, 435 – 440, 1990.

[2] Silverman H., *Complex Variables*, Houghton, Newark, USA, 1975.

[3] Leung H. and Haykin S., "The Complex Backpropagation Algorithm", *IEEE Trans. on Signal Processing*, Vol. 3, No. 9, pp. 2101 – 2104, 1991.

[4] Benvenuto N., Marchesi M., Piazza F., and Uncini A., "Non Linear Satellite Radio Links Equalized using Blind Neural Networks", *Proc. of ICASSP* Vol. 3, pp. 1521 –1524, 1991.

[5] Georgiou G. and Koutsougeras C, "Complex Backpropagation", *IEEE Trans. on Circuits and Systems II*, Vol. 39, No.5, May, pp. 330 – 334, 1992.

[6] Hirose A "Continuous Complex-Valued Back-propagation Learning", *Electronics Letters*, Vol.28, No. 20, pp. 1854 – 1855, September 24, 1992.

[7] Kim T. and Adali T. "Complex Backpropagtion Neural Network Using Elementary Transcendental Activation Functions", *Proc. of IEEE ICASSP*, *Proc.* Vol. II, Salt Lake City, May 2001.

[8] Hornik K., Stinchcombe M., and White H., "Multilayer Feedforward Networks are Universal Approximators", *Neural Networks*, Vol. 2, pp. 359-366, 1989

[9] Rudin W., "Real and Complex Analysis", TMH edition, McGraw Hill, Inc., 1974