Neural Networks 69 (2015) 44–50

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Convergence analysis of an augmented algorithm for fully complex-valued neural networks[☆]

Dongpo Xu[a,b,*], Huisheng Zhang[c], Danilo P. Mandic[d]

[a] School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, China
[b] College of Science, Harbin Engineering University, Harbin 150001, China
[c] Department of Mathematics, Dalian Maritime University, Dalian 116026, China
[d] Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ London, UK

ARTICLE INFO

Article history:
Received 29 June 2014
Received in revised form 28 February 2015
Accepted 17 May 2015
Available online 27 May 2015

Keywords:
Complex-valued neural networks
Augmented algorithm
Unified mean value theorem
Wirtinger calculus
Convergence

ABSTRACT

This paper presents an augmented algorithm for fully complex-valued neural network based on Wirtinger calculus, which simplifies the derivation of the algorithm and eliminates the Schwarz symmetry restriction on the activation functions. A unified mean value theorem is first established for general functions of complex variables, covering the analytic functions, non-analytic functions and real-valued functions. Based on so introduced theorem, convergence results of the augmented algorithm are obtained under mild conditions. Simulations are provided to support the analysis.

1. Introduction

Complex-valued neural networks (CVNNs) have recently attracted broad research interests, for example, in seismics, sonar, and radar (Hirose, 2012). CVNNs have been shown to inherent advantages in reducing the number of parameters and operations involved (Mandic & Goh, 2009). In addition, CVNNs have computational advantages over real-valued neural networks in solving classification problems (Aizenberg, 2011), and can even solve the XOR problem with only one complex-valued neuron (Nitta, 2003). However, the choice of activation function remains being a challenging task due to the conflict requirements of boundedness and analyticity—Liouville's theorem states that if a function is analytic and bounded in the complex plane, then it must be a constant. A traditional split-complex approach (Nitta, 1997) uses a pair of real-valued activation functions to process the real and imaginary parts of a complex signal separately. While this approach helps avoiding the problem of unboundedness, split complex activation functions are never analytic. In contrast, 'fully' complex activation functions (Kim & Adali, 2003), such as elementary transcendental functions, are analytic and bounded almost everywhere in $\mathbb{C}$, and have been used in multi-layer perceptions (Kim & Adali, 2003), radial basis function networks (Savitha, Suresh, & Sundararajan, 2012) and extreme learning machines (Li, Huang, Saratchandran, & Sundararajan, 2005). Classical real-valued learning algorithms that have been extended to complex case, contains the complex least mean square (Widrow, McCool, & Ball, 1975), complex backpropagation (Georgiou & Koutsougeras, 1992; Hirose, 1992; Leung & Haykin, 1991; Nitta, 1997) and complex real-time recurrent learning (Goh & Mandic, 2004, 2007a). Signal processing techniques (Adali, Li, Novey, & Cardoso, 2008; Dini & Mandic, 2012) have also been proposed based on such activation functions, however, the basic research issue: whether these complex algorithms share convergence properties with their real counterparts remains largely unanswered. The complex universal approximation theorem of the CVNNs with fully complex activation functions (denoted as FCVNNs for simplicity) has been given by Kim and Adali in Kim and Adali (2003), which ensures that the FCVNNs can be considered as a universal approximator of any continuous complex mappings.

Convergence of the real-valued learning algorithm has been widely studied (Shao & Zheng, 2011; Wang, Yang, & Wu, 2011; Wu, Fan, & Zurada, 2014; Wu, Feng, Li, & Xu, 2005; Wu, Wang, Cheng, & Li, 2011). However, in the complex domain, in addition to

[☆] This work was supported by the National Natural Science Foundation of China (Nos. 61301202, 61101228) and by the Research Fund for the Doctoral Program of Higher Education of China (No. 20122304120028).

[*] Corresponding author at: School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, China.

E-mail addresses: dongpoxu@gmail.com (D. Xu), zhhuisheng@163.com (H. Zhang), d.mandic@imperial.ac.uk (D.P. Mandic).

http://dx.doi.org/10.1016/j.neunet.2015.05.003
0893-6080/© 2015 Elsevier Ltd. All rights reserved.

the conflict between boundedness and analyticity of the activation function, another challenge is that the traditional mean value theorem does not hold in the complex domain (e.g., $f(z) = e^z$ with $z_2 = z_1 + 2\pi i$, then $f(z_1) = f(z_2)$ but $f(z_2) - f(z_1) \neq f'(\xi)(z_2 - z_1)$ for all $\xi \in \mathbb{C}$). In addition, cost functions are real-valued and therefore the complex derivative cannot be used. Some results for split-complex nonlinear gradient descent (SCNGD) algorithms exist (Xu, Zhang, & Liu, 2010; Zhang, Zhang, & Wu, 2009), whereby the analysis is based on reformulating complex algorithm in the real domain by separating it into real and imaginary parts. Furthermore, the convergence of the SCNGD algorithms with momentum or penalty has been established in Xu, Shao, and Zhang (2012) and Zhang, Xu, and Zhang (2014). In addition, the convergence of some complex adaptive filters algorithms has been obtained under the assumption that the activation function is a contraction (Mandic & Goh, 2009). The convergence of fully-complex nonlinear gradient descent (FCNGD) algorithms has been proved under Schwarz symmetry condition $f^*(z) = f(z^*)$ (Zhang, Liu, Xu, & Zhang, 2014). However, this condition is usually not valid for a polynomial function with complex coefficients, and the mean value theorem used in Zhang, Liu et al. (2014) is not applicable to the non-analytic functions, such as real-valued cost functions. Recently, augmented complex statistics have been introduced into some learning algorithms, such as the augmented complex least mean square (Mandic & Goh, 2009; Mandic, Javidi, Goh, Kuh, & Aihara, 2009), augmented complex extended Kalman filter (Dini & Mandic, 2012; Goh & Mandic, 2007b), and augmented echo state network (Xia, Jelfs, Van Hulle, Príncipe, & Mandic, 2011). These can capture the second-order statistical information and thus produce optimal estimates for second-order noncircular (improper) signals. However, the convergence of the augmented FCNGD (AFCNGD) algorithms for the FCVNNs has not yet been fully established in the literature, which motivates this work.

The aim of this paper is to present a comprehensive study on the weak and strong convergence for the AFCNGD algorithm, indicating that the gradient of the error function goes to zero and the weight sequence goes to a fixed point, respectively. In comparison to the existing complex backpropagation (CBP) algorithms (Georgiou & Koutsougeras, 1992; Hirose, 1992; Leung & Haykin, 1991; Nitta, 1997), the proposed AFCNGD algorithm shows faster convergence and better steady-state performance. The main points and novel contributions of this paper are as follows:

- Based on Wirtinger calculus, we develop an augmented FCNGD algorithm for CVNNs with fully complex activation functions. This approach can simplify the derivation of the proposed algorithm and eliminate Schwarz symmetry restriction on the complex activation functions.
- We establish a unified mean value theorem for the complex nonlinear functions, covering the analytic functions, non-analytic functions and real-valued functions. This theorem plays an important role in the convergence proof of the proposed AFCNGD algorithm.
- The deterministic convergence including weak convergence and strong convergence of the AFCNGD algorithm is obtained. Our results are of considerable generality, including as particular cases almost all CVNNs with complex elementary transcendental functions given in Kim and Adali (2003).
- Illustrated experiments have been performed to verify the theoretical results of this paper and the advantages of the proposed AFCNGD algorithm.

The rest of this paper is organized as follows. In Section 2, we provide an overview of second-order augmented complex statistics and Wirtinger calculus. Section 3 derives the proposed augmented learning algorithm for the FCVNNs. The main convergence results and their proofs are presented in Section 4. Supporting numerical experiments are presented in Section 5. Some conclusions are drawn in Section 6.

2. Preliminaries

2.1. Notations

We use bold-face upper case letter to denote matrices, bold-faced lower case letters for column vectors, and light-faced lower case letters for scalars. The superscripts $(\cdot)^*$, $(\cdot)^T$ and $(\cdot)^H$ denote the complex conjugate, transpose and Hermitian (conjugate transpose), respectively. $\mathrm{Re}(z)$ and $|z|$ denote the real part and module of a complex number z. $\|\mathbf{z}\|$ and $\|\mathbf{Z}\|$ denote the Frobenius norm of a vector $\mathbf{z}$ and a matrix $\mathbf{Z}$. Finally, we refer to $f(z^*) = f^*(z)$ as the Schwarz symmetry principle (Needham, 1998, p. 257).

2.2. Second-order augmented complex statistics

The recent introduction of so-called augmented complex statistics (Mandic & Goh, 2009) showed that for a general (improper) complex vector $\mathbf{z}$, second order statistics based on the covariance matrix $\mathbf{C}_{\mathbf{zz}} = E[\mathbf{z}\mathbf{z}^H]$ is inadequate and that the pseudo-covariance matrix $\mathbf{P}_{\mathbf{zz}} = E[\mathbf{z}\mathbf{z}^T]$ is also required to fully capture the second order information. Processes with the vanishing pseudo-covariance, $\mathbf{P}_{\mathbf{zz}} = \mathbf{0}$ is termed second order circular (or proper). In real-world applications, most complex signals are second order noncircular or improper, and their probability density functions are not rotation invariant. In practice, the widely linear modeling (Picinbono & Chevalier, 1995) is based on a regressor vector produced by concatenating the input vector $\mathbf{z}$ with its complex conjugate $\mathbf{z}^*$, to give an augmented $2M \times 1$ input vector $\mathbf{z}^a = [\mathbf{z}^T, \mathbf{z}^H]^T$, together with the corresponding augmented coefficient vector $\mathbf{w}^a = [\mathbf{u}^T, \mathbf{v}^H]^T$. The $2M \times 2M$ augmented covariance matrix (Schreier & Scharf, 2003) then becomes

$$\mathbf{C}_{\mathbf{z}^a\mathbf{z}^a} = E\begin{bmatrix}\mathbf{z}\\ \mathbf{z}^*\end{bmatrix}\begin{bmatrix}\mathbf{z}^H & \mathbf{z}^T\end{bmatrix} = \begin{pmatrix}\mathbf{C}_{\mathbf{zz}} & \mathbf{P}_{\mathbf{zz}}\\ \mathbf{P}^*_{\mathbf{zz}} & \mathbf{C}^*_{\mathbf{zz}}\end{pmatrix}. \tag{1}$$

This matrix now contains the complete complex second order statistical information available in the complex domain, see Mandic and Goh (2009) and Schreier and Scharf (2010) for more details.

2.3. Wirtinger calculus

Any function of a complex variable z can be defined as $f(z) = u(x, y) + iv(x, y)$, where $z = x + iy$ and i denotes an imaginary unit. If the partial derivatives $\frac{\partial u}{\partial y}$, $\frac{\partial v}{\partial x}$, $\frac{\partial u}{\partial x}$, $\frac{\partial v}{\partial y}$ exist and satisfy the Cauchy–Riemann conditions $\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}$ and $\frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}$, then $f(z)$ is said to be analytic (complex derivative exists), otherwise, it is non-analytic (complex derivative does not exist). For general functions of complex variables (both analytic and non-analytic), the following pair of derivatives can be defined (Brandwood, 1983; Kreutz-Delgado, 2009; Wirtinger, 1927)

$$\text{R-derivative: } \frac{\partial f}{\partial z} = \frac{1}{2}\left(\frac{\partial f}{\partial x} - i\frac{\partial f}{\partial y}\right) \tag{2}$$

$$\text{R}^*\text{-derivative: } \frac{\partial f}{\partial z^*} = \frac{1}{2}\left(\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial y}\right) \tag{3}$$

which are called Wirtinger or CR derivatives. In particular, if $f(z)$ is analytic, then the R-derivative $\frac{\partial f}{\partial z}$ becomes the complex derivative $f'(z)$ and the R*-derivative vanishes, that is the Cauchy–Riemann equations are equivalent to $\frac{\partial f}{\partial z^*} = 0$. Some basic rules of the CR derivatives are summarized as (Kreutz-Delgado, 2009; Mandic & Goh, 2009; Wirtinger, 1927)

$$\text{Differential rule: } df = \frac{\partial f}{\partial z}dz + \frac{\partial f}{\partial z^*}dz^* \tag{4}$$

$$\text{Chain rule: } \frac{\partial g(f)}{\partial z} = \frac{\partial g}{\partial f}\frac{\partial f}{\partial z} + \frac{\partial g}{\partial f^*}\frac{\partial f^*}{\partial z} \tag{5}$$

$$\text{Chain rule: } \frac{\partial g(f)}{\partial z^*} = \frac{\partial g}{\partial f}\frac{\partial f}{\partial z^*} + \frac{\partial g}{\partial f^*}\frac{\partial f^*}{\partial z^*} \tag{6}$$

$$\begin{aligned}&\text{Conjugation rule: } \left(\frac{\partial f}{\partial z}\right)^* = \frac{\partial f^*}{\partial z^*};\\ &\quad \text{when } f \text{ is real } \left(\frac{\partial f}{\partial z}\right)^* = \frac{\partial f}{\partial z^*}\end{aligned} \tag{7}$$

$$\begin{aligned}&\text{Conjugation rule: } \left(\frac{\partial f}{\partial z^*}\right)^* = \frac{\partial f^*}{\partial z};\\ &\quad \text{when } f \text{ is real } \left(\frac{\partial f}{\partial z^*}\right)^* = \frac{\partial f}{\partial z}.\end{aligned} \tag{8}$$

3. Learning algorithm for CVNN using Wirtinger calculus

For simplicity, we consider a three-layer FCVNN with M input nodes, M extended input nodes, N hidden nodes and one output node. Let $\mathbf{w}_0 = (w_{01}, w_{02}, \ldots, w_{0N})^T \in \mathbb{C}^N$ be the weighting vector between the hidden units and the output unit, and $\mathbf{u}_n, (\mathbf{v}_n) \in \mathbb{C}^M$ $(n = 1, 2, \ldots, N)$ be the weighting vectors between the input (extended input) units and the n-th hidden unit. All the network weights can be written in a compact form $\mathbf{w} = \left(\mathbf{w}_0^T, \mathbf{u}_1^T, \ldots, \mathbf{u}_N^T, \mathbf{v}_1^T, \ldots, \mathbf{v}_N^T\right)^T \in \mathbb{C}^{N(2M+1)}$. Let $f, g : \mathbb{C} \to \mathbb{C}$ be the fully complex activation functions of hidden and output layers. We now define two vector-valued functions

$$\begin{aligned}\mathbf{g}(\mathbf{z}) &= (g(z_1), g(z_2), \ldots, g(z_M))^T,\\ \mathbf{g}'(\mathbf{z}) &= (g'(z_1), g'(z_2), \ldots, g'(z_M))^T\end{aligned} \tag{9}$$

where $\mathbf{z} = (z_1, z_2, \ldots, z_M)^T \in \mathbb{C}^M$. For the input vector $\mathbf{z} \in \mathbb{C}^M$ and its conjugate $\mathbf{z}^*$, the output of the augmented FCVNN is given by

$$y = f\left(\mathbf{w}_0 \cdot \mathbf{g}(\mathbf{U}\mathbf{z} + \mathbf{V}\mathbf{z}^*)\right) \tag{10}$$

where '$\cdot$' denotes the inner product of two vectors, $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_N)^T \in \mathbb{C}^{N\times M}$ and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_N)^T \in \mathbb{C}^{N\times M}$. Note that the output of the FCVNN in (10) depends on both $\mathbf{z}$ and $\mathbf{z}^*$, so it is suitable for the processing of general complex valued signals, both circular and noncircular (Mandic & Goh, 2009).

For the training sample set $\{\mathbf{z}_j, d_j\}_{j=1}^J \subset \mathbb{C}^M \times \mathbb{C}$, where $\mathbf{z}_j$ and d_j are respectively the input and the desired output, the training process finds an optimal weighting vector $\mathbf{w}^\star$ that minimizes the error function

$$E(\mathbf{w}) = \sum_{j=1}^J |e_j|^2 = \sum_{j=1}^J |y_j - d_j|^2 \tag{11}$$

where

$$e_j = y_j - d_j, \quad y_j = f\left(\mathbf{w}_0 \cdot \mathbf{g}\left(\mathbf{U}\mathbf{z}_j + \mathbf{V}\mathbf{z}_j^*\right)\right). \tag{12}$$

Using the chain rule in (5) and noting that f, g are analytic (R*-derivative vanishes), we have

$$\begin{aligned}\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}_0} &= \sum_{j=1}^J \left(\frac{\partial |e_j|^2}{\partial e_j}\frac{\partial e_j}{\partial \mathbf{w}_0} + \frac{\partial |e_j|^2}{\partial e_j^*}\frac{\partial e_j^*}{\partial \mathbf{w}_0}\right)\\ &= \sum_{j=1}^J e_j^* \frac{\partial f\left(\mathbf{w}_0 \cdot \mathbf{g}\left(\mathbf{U}\mathbf{z}_j + \mathbf{V}\mathbf{z}_j^*\right)\right)}{\partial \mathbf{w}_0}\\ &\quad + \sum_{j=1}^J e_j \frac{\partial f^*\left(\mathbf{w}_0 \cdot \mathbf{g}\left(\mathbf{U}\mathbf{z}_j + \mathbf{V}\mathbf{z}_j^*\right)\right)}{\partial \mathbf{w}_0}\end{aligned} \tag{13}$$

while from the conjugate rule in (8) and $\frac{\partial z}{\partial z^*} = 0$, we arrive at

$$\begin{aligned}\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}_0} &= \sum_{j=1}^J e_j^* f'\left(\mathbf{w}_0 \cdot \mathbf{g}(\mathbf{U}\mathbf{z}_j + \mathbf{V}\mathbf{z}_j^*)\right) \mathbf{g}^T\left(\mathbf{U}\mathbf{z}_j + \mathbf{V}\mathbf{z}_j^*\right)\\ &\quad + \sum_{j=1}^J e_j \Bigg(f'\left(\mathbf{w}_0 \cdot \mathbf{g}(\mathbf{U}\mathbf{z}_j + \mathbf{V}\mathbf{z}_j^*)\right)\\ &\quad \times \frac{\partial\left(\mathbf{w}_0 \cdot \mathbf{g}(\mathbf{U}\mathbf{z}_j + \mathbf{V}\mathbf{z}_j^*)\right)}{\partial \mathbf{w}_0^*}\Bigg)^*\\ &= \sum_{j=1}^J e_j^* f'\left(\mathbf{w}_0 \cdot \mathbf{g}(\mathbf{U}\mathbf{z}_j + \mathbf{V}\mathbf{z}_j^*)\right) \mathbf{g}^T\left(\mathbf{U}\mathbf{z}_j + \mathbf{V}\mathbf{z}_j^*\right).\end{aligned} \tag{14}$$

Note that this derivation does not required the Schwarz symmetry $f^*(z) = f(z^*)$ that is needed in Adali et al. (2008), Leung and Haykin (1991) and Zhang, Liu et al. (2014). Similarly, we have

$$\begin{aligned}\frac{\partial E(\mathbf{w})}{\partial \mathbf{u}_n} &= \sum_{j=1}^J e_j^* f'\left(\mathbf{w}_0 \cdot \mathbf{g}(\mathbf{U}\mathbf{z}_j + \mathbf{V}\mathbf{z}_j^*)\right)\\ &\quad \times w_{0n} g'\left(\mathbf{u}_n^T\mathbf{z}_j + \mathbf{v}_n^T\mathbf{z}_j^*\right)\mathbf{z}_j^T\end{aligned} \tag{15}$$

$$\begin{aligned}\frac{\partial E(\mathbf{w})}{\partial \mathbf{v}_n} &= \sum_{j=1}^J e_j^* f'\left(\mathbf{w}_0 \cdot \mathbf{g}(\mathbf{U}\mathbf{z}_j + \mathbf{V}\mathbf{z}_j^*)\right)\\ &\quad \times w_{0n} g'\left(\mathbf{u}_n^T\mathbf{z}_j + \mathbf{v}_n^T\mathbf{z}_j^*\right)\mathbf{z}_j^H.\end{aligned} \tag{16}$$

Since $E(\mathbf{w})$ is real-valued, then using the conjugation rule in (7), the gradient of $E(\mathbf{w})$ is given by

$$\begin{aligned}\nabla_{\mathbf{w}^*} E(\mathbf{w}) &\triangleq \left(\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}}\right)^H\\ &= \Bigg(\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}_0}, \frac{\partial E(\mathbf{w})}{\partial \mathbf{u}_1}, \ldots, \frac{\partial E(\mathbf{w})}{\partial \mathbf{u}_N},\\ &\qquad \frac{\partial E(\mathbf{w})}{\partial \mathbf{v}_1}, \ldots, \frac{\partial E(\mathbf{w})}{\partial \mathbf{v}_N}\Bigg)^H\end{aligned} \tag{17}$$

which defines the direction of the maximum rate of change of $E(\mathbf{w})$ in Brandwood (1983). Starting from an arbitrary initial weighting vector $\mathbf{w}^0$, the weight updating rule based on the Wirtinger calculus can be written as

$$\mathbf{w}^{k+1} = \mathbf{w}^k + \Delta\mathbf{w}^k = \mathbf{w}^k - \eta\nabla_{\mathbf{w}^*} E(\mathbf{w}^k) \tag{18}$$

where $\eta > 0$ is the learning rate. This completes the derivation of the augmented FCNGD (AFCNGD) algorithm.

4. Convergence analysis

To analyze the convergence of the algorithm (18), we need the following assumptions:

(A1) There exists a constant c_1 such that $\max\left\{\|\mathbf{w}_0^k\|, \|\mathbf{U}^k\|, \|\mathbf{V}^k\|\right\} \le c_1$ for all $k = 0, 1, 2, \ldots$;
(A2) The functions $f(z)$ and $g(z)$ are analytic in a bounded region $|z| < R$ and continuous on $|z| = R$, where R is defined in (25);
(A3) There exists a bounded closed region $\Phi \subset \mathbb{C}^{N(2M+1)}$ such that $\mathbf{w}_k \in \Phi$ and $\Phi_0 = \{\mathbf{w} \in \Phi : \nabla_{\mathbf{w}^*} E(\mathbf{w}) = \mathbf{0}\}$ contains only finite number of points.

Remark 4.1. Assumption (A1) means the condition on boundedness of $\|\mathbf{w}^k\|$, which is often used in the literature (Aizenberg, 2010, 2011; Gori & Maggini, 1996; Shao & Zheng, 2011; Wu et al., 2005, 2011; Xu et al., 2010), and can be removed when adding a penalty term to the error function (Zhang, Xu et al., 2014). Assumption (A2) indicates that complex coefficient polynomials can be used as the activation function, which removes the Schwarz symmetry condition in Adali et al. (2008), Leung and Haykin (1991) and Zhang, Liu et al. (2014). Assumption (A3) implies that the error function has only a finite number of local minima, which will be used to obtain strong convergence results.

Theorem 4.1 (*Mean Value Theorem of Integral Form*). *Consider a continuous function $f : S \subseteq \mathbb{C} \to \mathbb{C}$ for which the Wirtinger derivatives exist and are continuous in the set S. If $\exists z_1, z_0 \in S$ such that the segment joining them is also in S, then*

$$f(z_1) - f(z_0) = \int_0^1 \left(\frac{\partial}{\partial z}\Delta z + \frac{\partial}{\partial z^*}(\Delta z)^*\right) f(z_0 + t\Delta z)dt \tag{19}$$

where $\Delta z = z_1 - z_0$, and $\frac{\partial f}{\partial z}$, $\frac{\partial f}{\partial z^*}$ are the Wirtinger derivatives.

Proof. Denote $g(t) = f(z_0 + t\Delta z)$, $0 \le t \le 1$, then $g(t)$ is continuous on $[0, 1]$ and has real derivatives in $(0, 1)$. Using the chain rule (5), the derivative of $g(t)$ can be found as

$$g'(t) = \frac{\partial f(z_0 + t\Delta z)}{\partial z}\Delta z + \frac{\partial f(z_0 + t\Delta z)}{\partial z^*}(\Delta z)^*. \tag{20}$$

Upon substituting (20) into $g(1) - g(0) = \int_0^1 g'(t)dt$ with $g(0) = f(z_0)$ and $g(1) = f(z_1)$, the equality (19) follows. □

In particular, if $f(z)$ is analytic (R*-derivative vanishes), then (19) becomes

$$f(z_1) - f(z_0) = \int_0^1 f'(z_0 + t\Delta z)\Delta z dt. \tag{21}$$

Moreover, if $f(z)$ is real-valued, then

$$f(z_1) - f(z_0) = 2\int_0^1 \mathrm{Re}\left(\frac{\partial f(z_0 + t\Delta z)}{\partial z}\Delta z\right) dt \tag{22}$$

where the conjugation rule (7) is used in (22) and $\mathrm{Re}(z)$ is the real part of z.

Lemma 4.2. *If the assumptions* (A1) *and* (A2) *are valid. Then the gradient $\nabla_{\mathbf{w}^*}E(\mathbf{w})$ satisfies the Lipschitz condition, in other words, there exists a positive constant L, such that*

$$\|\nabla_{\mathbf{w}^*}E(\mathbf{w}^{k+1}) - \nabla_{\mathbf{w}^*}E(\mathbf{w}^k)\| \le L\|\mathbf{w}^{k+1} - \mathbf{w}^k\|. \tag{23}$$

Proof. For brevity, we use the following notion

$$\mathbf{s}^{k,j} = (\mathbf{U}^k)^T\mathbf{z}_j + (\mathbf{V}^k)^T\mathbf{z}_j^*, \qquad p^{k,j} = \mathbf{w}_0^k \cdot \mathbf{g}\left(\mathbf{s}^{k,j}\right) \tag{24}$$

where $k = 1, 2, \dots$; $j = 1, 2, \dots, J$. By Assumption (A1) and a finite set of samples $\{\mathbf{z}_j, d_j\}_{j=1}^J$, there exists a constant R such that

$$\max\left\{\left\|\mathbf{s}^{k,j}\right\|, \left|p^{k,j}\right|\right\} < R, \quad j = 1, 2, \dots, J;\ k = 1, 2, \dots. \tag{25}$$

Since the continuous functions f and g are bounded in the closed region $|z| \le R$, there exists $c_2 \in \mathbb{R}^+$ such that

$$\max\left\{|f(z)|, |f'(z)|, |f''(z)|, \|\mathbf{g}(\mathbf{z})\|, \left\|\mathbf{g}'(\mathbf{z})\right\|\right\} < c_2 \tag{26}$$

where the analyticity of f and g guarantees derivatives of all orders. By (21), (26) and the Cauchy–Schwarz inequality, we have

$$\begin{aligned}\left\|\mathbf{g}\left(\mathbf{s}^{k+1,j}\right) - \mathbf{g}\left(\mathbf{s}^{k,j}\right)\right\| &= \left\|\begin{matrix} g\left(s_1^{k+1,j}\right) - g\left(s_1^{k,j}\right) \\ \vdots \\ g\left(s_N^{k+1,j}\right) - g\left(s_N^{k,j}\right)\end{matrix}\right\| \\ &= \left\|\begin{matrix}\left(s_1^{k+1,j} - s_1^{k,j}\right)\int_0^1 g'(s_1^{k,j} + t\Delta s_1)dt \\ \vdots \\ \left(s_N^{k+1,j} - s_N^{k,j}\right)\int_0^1 g'(s_N^{k,j} + t\Delta s_N)dt\end{matrix}\right\| \\ &\le c_3\left(\|\mathbf{U}^{k+1} - \mathbf{U}^k\| + \|\mathbf{V}^{k+1} - \mathbf{V}^k\|\right), \end{aligned}\tag{27}$$

where $c_3 = c_2 \max_{1\le j\le J}\|\mathbf{z}_j\|$ and $\Delta s_n = s_n^{k+1,j} - s_n^{k,j}$. By (26), (27) and Assumption (A1), we now have

$$\begin{aligned}\left\|p^{k+1,j} - p^{k,j}\right\| &= \left\|\mathbf{w}_0^{k+1} \cdot \mathbf{g}\left(\mathbf{s}^{k+1,j}\right) - \mathbf{w}_0^k \cdot \mathbf{g}\left(\mathbf{s}^{k,j}\right)\right\| \\ &\le \left\|\left(\mathbf{w}_0^{k+1} - \mathbf{w}_0^k\right) \cdot \mathbf{g}\left(\mathbf{s}^{k+1,j}\right)\right\| \\ &\quad + \left\|\mathbf{w}_0^k \cdot \left(\mathbf{g}(\mathbf{s}^{k+1,j}) - \mathbf{g}(\mathbf{s}^{k,j})\right)\right\| \\ &\le c_2\left\|\mathbf{w}_0^{k+1} - \mathbf{w}_0^k\right\| + c_1\left\|\mathbf{g}\left(\mathbf{s}^{k+1,j}\right) - \mathbf{g}\left(\mathbf{s}^{k,j}\right)\right\| \\ &\le c_2\left\|\mathbf{w}_0^{k+1} - \mathbf{w}_0^k\right\| \\ &\quad + c_1c_3\left(\|\mathbf{U}^{k+1} - \mathbf{U}^k\| + \|\mathbf{V}^{k+1} - \mathbf{V}^k\|\right). \end{aligned}\tag{28}$$

Note that $\|\cdot\|$ is Frobenius norm, $\mathbf{w} = \left(\mathbf{w}_0^T, \mathbf{u}_1^T, \dots, \mathbf{u}_N^T, \mathbf{v}_1^T, \dots, \mathbf{v}_N^T\right)^T$, $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_N)$ and $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_N)$, which yields

$$\begin{aligned}\left\|p^{k+1,j} - p^{k,j}\right\| &\le c_2\left\|\mathbf{w}_0^{k+1} - \mathbf{w}_0^k\right\| \\ &\quad + c_1c_3\left(\|\mathbf{U}^{k+1} - \mathbf{U}^k\| + \|\mathbf{V}^{k+1} - \mathbf{V}^k\|\right) \\ &\le c_4\|\mathbf{w}^{k+1} - \mathbf{w}^k\| \end{aligned}\tag{29}$$

where $c_4 = 2\max\{c_2, c_1c_3\}$. Upon combining (21) with (29), this yields

$$\begin{aligned}\left\|y_j^{k+1} - y_j^k\right\| &= \left\|f(p^{k+1,j}) - f(p^{k,j})\right\| \\ &= \left\|\left(p^{k+1,j} - p^{k,j}\right)\int_0^1 f'(p^{k,j} + t\Delta p)dt\right\| \\ &\le c_2\left\|p^{k+1,j} - p^{k,j}\right\| \le c_5\|\mathbf{w}^{k+1} - \mathbf{w}^k\| \end{aligned}\tag{30}$$

where $c_5 = c_2c_4$. In the same way, we can prove that

$$\left\|f'(p^{k+1,j}) - f'(p^{k,j})\right\| \le c_5\|\mathbf{w}^{k+1} - \mathbf{w}^k\|. \tag{31}$$

From (21), (26)–(31) and the Cauchy–Schwarz inequality, we then have

$$\begin{aligned}&\|f'(p^{k+1,j})\mathbf{g}^T(\mathbf{s}^{k+1,j}) - f'(p^{k,j})\mathbf{g}^T(\mathbf{s}^{k,j})\| \\ &\quad \le \left\|\left(f'(p^{k+1,j}) - f'(p^{k,j})\right)\mathbf{g}^T\left(\mathbf{s}^{k+1,j}\right)\right\| \\ &\qquad + \left\|f'(p^{k,j})\left(\mathbf{g}(\mathbf{s}^{k+1,j}) - \mathbf{g}(\mathbf{s}^{k,j})\right)^T\right\| \\ &\quad \le c_5\|\mathbf{w}^{k+1} - \mathbf{w}^k\|\left\|\mathbf{g}\left(\mathbf{s}^{k+1,j}\right)\right\| \\ &\qquad + \left\|f'(p^{k,j})\right\|\left\|\mathbf{g}(\mathbf{s}^{k+1,j}) - \mathbf{g}(\mathbf{s}^{k,j})\right\| \\ &\quad \le c_2c_5\|\mathbf{w}^{k+1} - \mathbf{w}^k\| + c_2c_3\left(\|\mathbf{U}^{k+1} - \mathbf{U}^k\| + \|\mathbf{V}^{k+1} - \mathbf{V}^k\|\right) \\ &\quad \le c_2(c_3 + c_5)\|\mathbf{w}^{k+1} - \mathbf{w}^k\| \end{aligned}\tag{32}$$

while from (26), (30), (32) and the Cauchy–Schwarz inequality, we arrive at

$$\begin{aligned}&\left\|\frac{\partial E(\mathbf{w}^{k+1})}{\partial \mathbf{w}_0} - \frac{\partial E(\mathbf{w}^k)}{\partial \mathbf{w}_0}\right\| \\ &\quad = \left\|\sum_{j=1}^J (e_j^{k+1})^* f'(p^{k+1,j})\mathbf{g}^T\left(\mathbf{s}^{k+1,j}\right) - \sum_{j=1}^J (e_j^k)^* f'(p^{k,j})\mathbf{g}^T\left(\mathbf{s}^{k,j}\right)\right\| \\ &\quad \le \sum_{j=1}^J \left\|(y_j^{k+1} - y_j^k)^* f'(p^{k+1,j})\mathbf{g}^T\left(\mathbf{s}^{k+1,j}\right)\right\| \end{aligned}$$

$$
\begin{aligned}
&+\sum_{j=1}^{J}\left\|(e_j^k)^*\left(f'(p^{k+1,j})\mathbf{g}^T(\mathbf{s}^{k+1,j})-f'(p^{k,j})\mathbf{g}^T(\mathbf{s}^{k,j})\right)\right\|\\
&\le Jc_2^2c_5\|\mathbf{w}^{k+1}-\mathbf{w}^k\|+J\max_{j,k}|e_j^k|c_2(c_3+c_5)\|\mathbf{w}^{k+1}-\mathbf{w}^k\|\\
&\le L_1\|\mathbf{w}^{k+1}-\mathbf{w}^k\| \qquad (33)
\end{aligned}
$$

where $L_1 = Jc_2^2c_5 + J\max_{j,k}|e_j^k|c_2(c_3+c_5)$. Similarly,

$$
\begin{aligned}
\left\|\frac{\partial E(\mathbf{w}^{k+1})}{\partial \mathbf{u}_n}-\frac{\partial E(\mathbf{w}^{k})}{\partial \mathbf{u}_n}\right\| &= \left\|\frac{\partial E(\mathbf{w}^{k+1})}{\partial \mathbf{v}_n}-\frac{\partial E(\mathbf{w}^{k})}{\partial \mathbf{v}_n}\right\|\\
&\le L_2\|\mathbf{w}^{k+1}-\mathbf{w}^k\|. \qquad (34)
\end{aligned}
$$

Hence, (23) follows by setting $L = 2\max\{L_1, L_2\}$. □

The following lemma is a complex version of Theorem 14.1.5 in Ortega and Rheinboldt (1970), its proof is omitted as it follows straightforwardly.

Lemma 4.3. *Let $F : \Phi \subset \mathbb{C}^m \to \mathbb{C}^m$ $(m \ge 1)$ be continuous for a bounded closed region Φ. If the set $\Phi_0 = \{\mathbf{z} \in \Phi : F(\mathbf{z}) = 0\}$ has a finite number of points and the sequence $\{\mathbf{z}_k\} \subset \Phi$ satisfies:*

(1) $\lim_{k\to\infty} F(\mathbf{z}_k) = 0$,
(2) $\lim_{k\to\infty} \|\mathbf{z}_{k+1} - \mathbf{z}_k\| = 0$,

then there exists $\mathbf{z}^\star \in \Phi_0$ such that $\lim_{k\to\infty} \mathbf{z}_k = \mathbf{z}^\star$.

Theorem 4.4. *Let the sequence $\{\mathbf{w}^k, k = 1, 2, \ldots\}$ be generated by the algorithm (18) with an initial value $\mathbf{w}^0$. If Assumptions (A1) and (A2) are valid and the learning rate η satisfies (38), then the conclusion for weak convergence is given by*

$$\lim_{k\to+\infty}\left\|\nabla_{\mathbf{w}^*}E(\mathbf{w}^k)\right\| = 0. \qquad (35)$$

Furthermore, if Assumption (A3) is valid, the conclusion for strong convergence states that there exists $\mathbf{w}^\star \in \Phi_0$ such that

$$\lim_{k\to+\infty} \mathbf{w}^k = \mathbf{w}^\star. \qquad (36)$$

Proof. By (22), Lemma 4.2 and Cauchy–Schwarz inequality, we have

$$
\begin{aligned}
E(\mathbf{w}^{k+1}) - E(\mathbf{w}^k) &= 2\int_0^1 \mathrm{Re}\left(\frac{\partial E(\mathbf{w}^k + t\Delta\mathbf{w}^k)}{\partial \mathbf{w}}\Delta\mathbf{w}^k\right)dt\\
&= 2\int_0^1 \mathrm{Re}\left((\nabla_{\mathbf{w}^*}E(\mathbf{w}^k + t\Delta\mathbf{w}^k))^H\Delta\mathbf{w}^k\right)dt\\
&= 2\mathrm{Re}\left((\nabla_{\mathbf{w}^*}E(\mathbf{w}^k))^H\Delta\mathbf{w}^k\right)\\
&\quad + 2\int_0^1 \mathrm{Re}\left((\nabla_{\mathbf{w}^*}E(\mathbf{w}^k + t\Delta\mathbf{w}^k) - \nabla_{\mathbf{w}^*}E(\mathbf{w}^k))^H\Delta\mathbf{w}^k\right)dt\\
&\le 2\mathrm{Re}\left((\nabla_{\mathbf{w}^*}E(\mathbf{w}^k))^H\Delta\mathbf{w}^k\right)\\
&\quad + 2\int_0^1 \left\|\nabla_{\mathbf{w}^*}E(\mathbf{w}^k + t\Delta\mathbf{w}^k) - \nabla_{\mathbf{w}^*}E(\mathbf{w}^k)\right\|\ \|\Delta\mathbf{w}^k\|dt\\
&\le -2\eta\mathrm{Re}\left((\nabla_{\mathbf{w}^*}E(\mathbf{w}^k))^H\nabla_{\mathbf{w}^*}E(\mathbf{w}^k)\right) + 2L\int_0^1 t\|\Delta\mathbf{w}^k\|^2dt\\
&= \eta(-2+\eta L)\left\|\nabla_{\mathbf{w}^*}E(\mathbf{w}^k)\right\|^2. \qquad (37)
\end{aligned}
$$

Further, if the learning rate η is small enough such that

$$0 < \eta < \frac{2}{L} \qquad (38)$$

then $E(\mathbf{w}^{k+1}) \le E(\mathbf{w}^k)$ holds. Write $\beta = \eta(2-\eta L)$. According to (37), it suffices to show that

$$
\begin{aligned}
E(\mathbf{w}^{k+1}) &\le E(\mathbf{w}^k) - \beta\left\|\nabla_{\mathbf{w}^*}E(\mathbf{w}^k)\right\|^2\\
&\le \cdots \le E(\mathbf{w}^0) - \beta\sum_{n=0}^{k}\left\|\nabla_{\mathbf{w}^*}E(\mathbf{w}^n)\right\|^2. \qquad (39)
\end{aligned}
$$

Since $E(\mathbf{w}^{k+1}) \ge 0$, for $k \to +\infty$ we obtain

$$\beta\sum_{n=0}^{+\infty}\left\|\nabla_{\mathbf{w}^*}E(\mathbf{w}^n)\right\|^2 \le E(\mathbf{w}^0). \qquad (40)$$

This immediately gives

$$\lim_{k\to+\infty}\left\|\nabla_{\mathbf{w}^*}E(\mathbf{w}^k)\right\| = 0 \qquad (41)$$

which completes the proof of weak convergence.

Next, to prove the strong convergence, from (18) and (41), we have

$$\lim_{k\to+\infty}\left\|\mathbf{w}^{k+1} - \mathbf{w}^k\right\| = \eta\lim_{k\to+\infty}\left\|\nabla_{\mathbf{w}^*}E(\mathbf{w}^k)\right\| = 0. \qquad (42)$$

Using Lemma 4.3 and by taking $\mathbf{z} = \mathbf{w}$ and $F(\mathbf{z}) = \nabla_{\mathbf{w}^*}E(\mathbf{w})$, together with Assumption (A3) and (42), this immediately leads to the strong convergence, i.e., there exists $\mathbf{w}^\star \in \Phi_0$ such that $\lim_{k\to\infty}\mathbf{w}^k = \mathbf{w}^\star$. □

5. Simulations

In the simulations, almost all the complex elementary transcendental functions can be selected as the activation function within the FCVNNs. However, for comparison purposes, the nonlinearities within neurons of the CVNN were chosen to be the following fully-complex tanh function (Leung & Haykin, 1991), split-complex tanh function (Nitta, 1997) and amplitude–phase functions (Georgiou & Koutsougeras, 1992; Hirose, 1992)

$$
\begin{aligned}
&\Phi_{fc}(z) = \tanh(z), \quad \Phi_{sc}(z) = \tanh(x) + i\tanh(y)\\
&\Phi_{ap}(z) = \tanh(|z|)\exp(i\arg z), \qquad \Phi_{ap}(z) = \frac{z}{c + \frac{1}{r}|z|} \qquad (43)
\end{aligned}
$$

where $z = x + iy \in \mathbb{C}$ $(x, y$ real numbers$)$, $c = 1$ and $r = 1$. The FCVNN architecture consisted of one output node and $N = 6$ hidden nodes, with the tap input length of $M = 2$. The randomly selected initial weights $\mathbf{w}^0$ were taken from a uniform distribution in the range $[-1, +1]$, the learning rate $\eta = 0.001$, and the training procedure was epochwise, with 1000 epochs of $J = 100$ training samples.

In the first set of experiments, we illustrate the convergence behavior of the AFCNGD algorithm by averaging the performance curves of 10 independent trials for one-step-ahead prediction of complex-valued signals. The complex benchmark noncircular signal was a complex AR moving-average (ARMA) process, given by (Xia et al., 2011)

$$
\begin{aligned}
z(t) = {}& 1.79z(t-1) - 1.85z(t-2) + 1.27z(t-3)\\
&- 0.41z(t-4) + 0.2z(t-5) + 2n(t) + 0.5n^*(t)\\
&+ n(t-1) + 0.9n^*(t-1) \qquad (44)
\end{aligned}
$$

where $n(t)$ is the complex-valued doubly white circular Gaussian noise, while the complex benchmark nonlinear input signal is given by (Narendra & Parthasarathy, 1990)

$$z(t) = \frac{z(t-1)}{1 + z^2(t-1)} + n^3(t). \qquad (45)$$

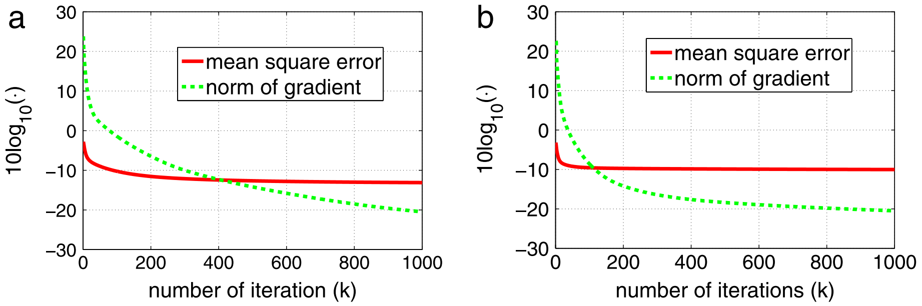

Fig. 1. Learning curves of AFCNGD for the noncircular input (44) and nonlinear input (45) for $\eta = 0.001$. (a) Noncircular signal. (b) Nonlinear signal.

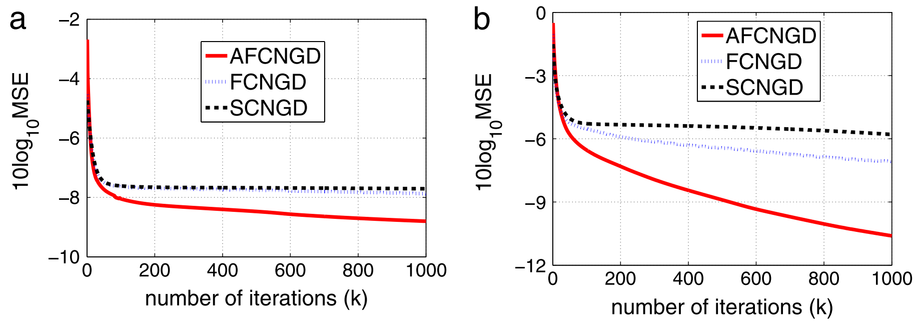

Fig. 2. Learning curves of AFCNGD, FCNGD and SCNGD for the nonlinear signal (44) and Ikeda map signal (46) for $\eta = 0.001$. (a) Nonlinear signal. (b) Ikeda map signal.

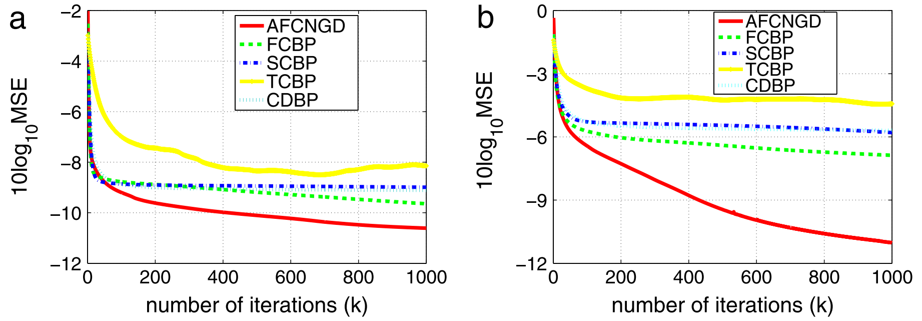

Fig. 3. Comparison of five kinds of algorithms for the nonlinear signal (44) and Ikeda map signal (46) for $\eta = 0.001$. (a) Nonlinear signal. (b) Ikeda map signal.

Fig. 1 shows the learning curves for the AFCNGD algorithm on complex noncircular (44) and nonlinear (45) signals, indicating that the mean square error (MSE) decreases monotonically and correspondingly the gradient converges to zero in magnitude along the iterations. Thus, the simulation results support our convergence theorem in Section 4.

In the second set of simulations, we compared the performance of AFCNGD with standard FCNGD and SCNGD (without the augmented states) on the prediction of the nonlinear signal (45) and synthetic nonlinear and noncircular chaotic Ikeda map signal (Aihara, 1994), given by

$$\begin{aligned} x(t+1) &= 1 + \mu[x(t)\cos(\alpha(t)) - y(t)\sin(\alpha(t))] \\ y(t+1) &= \mu[x(t)\sin(\alpha(t)) + y(t)\cos(\alpha(t))] \end{aligned} \tag{46}$$

where $\mu = 0.9$ and $\alpha(t) = 0.4 - 6/(1 + x^2(t) + y^2(t))$.

Fig. 2 shows the prediction performance of the AFCNGD applied to the complex-valued nonlinear signal (45) and chaotic Ikeda map signal (46). In both cases, there was a significant improvement in the performance when the AFCNGD was employed over that of the FCNGD and SCNGD algorithms.

To further illustrate the advantage of the AFCNGD using the augmented complex statistics, four complex backpropagation algorithms, namely, teachers-signal complex backpropagation (TCBP) (Hirose, 1992), complex domain backpropagation (CDBP) (Georgiou & Koutsougeras, 1992), fully-complex backpropagation (FCBP) (Leung & Haykin, 1991), and split-complex backpropagation (SCBP) (Nitta, 1997) algorithm are provided. These four kinds of algorithms are compared to the proposed AFCNGD algorithm for nonlinear signals (45) and chaotic Ikeda map signal (46). In Fig. 3, we give the learning curves of these five algorithms for the two types of complex signal. We can see that the AFCNGD algorithm converges quickly, and the steady state performances of the AFCNGD algorithm is better than those of the CBP algorithms. From this comparison, we note that the augmented complex statics provide an increment of performance, which is consistent with the theoretical results shown in Picinbono and Chevalier (1995).

6. Conclusions

The AFCNGD algorithm has been introduced for training FCVNNs under the framework of Wirtinger calculus, which greatly reduces the algorithm derivation and removes the Schwarz symmetry restriction on the complex activation functions within FCVNNs. Further, a unified mean value theorem has been introduced for general functions of complex variables, both analytic and non-analytic ones. This has enabled us to prove both the weak and strong convergence results of the proposed AFCNGD algorithm. The results so obtained are valid for more extensive classes of CVNNs, including the CVNNs with complex hyperbolic tangent activation functions as a special case. Illustrative experiments are implemented to illustrate theoretical results, and the comparison between the AFCNGD algorithm and the existing CBP algorithms shows that augmented complex statistics plays an important role in improving the convergence speed and steady state performance.

Acknowledgments

The authors wish to thank the anonymous reviewers for their insightful comments and suggestions which greatly improved this work.

References

Adali, T., Li, H., Novey, M., & Cardoso, J. (2008). Complex ICA using nonlinear functions. *IEEE Transactions on Signal Processing*, *56*, 4536–4544.

Aihara, K. (1994). *Applied chaos and applicable chaos*. Tokyo, Japan: Science-Sha.

Aizenberg, I. (2010). Periodic activation function and a modified learning algorithm for the multivalued neuron. *IEEE Transactions on Neural Networks*, *21*, 1939–1949.

Aizenberg, I. (2011). *Complex-valued neural networks with multivalued neurons*. Heidelberg: Springer.

Brandwood, D. (1983). A complex gradient operator and its application in adaptive array theory. *IEEE Communications, Radar and Signal Processing*, *130*, 11–16.

Dini, D. H., & Mandic, D. P. (2012). Class of widely linear complex Kalman filters. *IEEE Transactions on Neural Networks and Learning Systems*, *23*, 775–786.

Georgiou, G. M., & Koutsougeras, C. (1992). Complex domain backpropagation. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, *39*, 330–334.

Goh, S. L., & Mandic, D. P. (2004). A complex-valued RTRL algorithm for recurrent neural networks. *Neural Computation*, *16*, 2699–2713.

Goh, S. L., & Mandic, D. P. (2007a). An augmented CRTRL for complex-valued recurrent neural networks. *Neural Networks*, *20*, 1061–1066.

Goh, S. L., & Mandic, D. P. (2007b). An augmented extended Kalman filter algorithm for complex-valued recurrent neural networks. *Neural Computation*, *19*, 1039–1055.

Gori, M., & Maggini, M. (1996). Optimal convergence of on-line backpropagation. *IEEE Transactions on Neural Networks*, *17*, 251–254.

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

Hirose, A. (2012). *Complex-valued neural networks* (2nd ed.). New York: Springer.

Kim, T., & Adali, T. (2003). Approximation by fully complex multilayer perceptrons. *Neural Computation*, *15*, 1641–1666.

Kreutz-Delgado, K. (2009). The complex gradient operator and the CR-calculus (pp. 1–74). ArXiv Preprint arXiv:0906.4835.

Leung, H., & Haykin, S. (1991). The complex backpropagation algorithm. *IEEE Transactions on Signal Processing*, *39*, 2101–2104.

Li, M., Huang, G., Saratchandran, P., & Sundararajan, N. (2005). Fully complex extreme learning machine. *Neurocomputing*, *68*, 306–314.

Mandic, D. P., & Goh, S. L. (2009). *Complex valued nonlinear adaptive filters: noncircularity, widely linear and neural models*. Wiley.

Mandic, D. P., Javidi, S., Goh, S. L., Kuh, A., & Aihara, K. (2009). Complex-valued prediction of wind profile using augmented complex statistics. *Renewable Energy*, *34*, 196–201.

Narendra, K. S., & Parthasarathy, K. (1990). Identification and control of dynamical systems using neural networks. *IEEE Transactions on Neural Networks*, *1*, 4–27.

Needham, T. (1998). *Visual complex analysis*. Oxford University Press.

Nitta, T. (1997). An extension of the back-propagation algorithm to complex numbers. *Neural Networks*, *10*, 1391–1415.

Nitta, T. (2003). Solving the XOR problem and the detection of symmetry using a single complex-valued neuron. *Neural Networks*, *16*, 1101–1105.

Ortega, J. M., & Rheinboldt, W. C. (1970). *Iterative solution of nonlinear equations in several variables*. New York: Academic Press.

Picinbono, B., & Chevalier, P. (1995). Widely linear estimation with complex data. *IEEE Transactions on Signal Processing*, *43*, 2030–2033.

Savitha, R., Suresh, S., & Sundararajan, N. (2012). Metacognitive learning in a fully complex-valued radial basis function neural network. *Neural Computation*, *24*, 1297–1328.

Schreier, P. J., & Scharf, L. L. (2003). Second-order analysis of improper complex random vectors and process. *IEEE Transactions on Signal Processing*, *51*, 714–725.

Schreier, P. J., & Scharf, L. L. (2010). *Statistical signal processing of complex-valued data: the theory of improper and noncircular signals*. Cambridge University Press.

Shao, H., & Zheng, G. (2011). Convergence analysis of a back-propagation algorithm with adaptive momentum. *Neurocomputing*, *74*, 749–752.

Wang, J., Yang, J., & Wu, W. (2011). Convergence of cyclic and almost-cyclic learning with momentum for feedforward neural networks. *IEEE Transactions on Neural Networks*, *22*, 1297–1306.

Widrow, B., McCool, J., & Ball, M. (1975). The complex LMS algorithm. *Proceedings of the IEEE*, *63*, 712–720.

Wirtinger, W. (1927). Zur formalen theorie der funktionen von mehr komplexen veränderlichen. *Mathematische Annalen*, *97*, 357–375.

Wu, W., Fan, Q., Zurada, J. M., et al. (2014). Batch gradient method with smoothing $L1/2$ regularization for training of feedforward neural networks. *Neural Networks*, *50*, 72–78.

Wu, W., Feng, G., Li, Z., & Xu, Y. (2005). Deterministic convergence of an online gradient method for BP neural networks. *IEEE Transactions on Neural Networks*, *16*, 533–540.

Wu, W., Wang, J., Cheng, M., & Li, Z. (2011). Convergence analysis of online gradient method for BP neural networks. *Neural Networks*, *24*, 91–98.

Xia, Y., Jelfs, B., Van Hulle, M. M., Príncipe, J. C., & Mandic, D. P. (2011). An augmented echo state network for nonlinear adaptive filtering of complex noncircular signals. *IEEE Transactions on Neural Networks*, *22*, 74–83.

Xu, D., Shao, H., & Zhang, H. (2012). A new adaptive momentum algorithm for split-complex recurrent neural networks. *Neurocomputing*, *93*, 133–136.

Xu, D., Zhang, H., & Liu, L. (2010). Convergence analysis of three classes of split-complex gradient algorithms for complex-valued recurrent neural networks. *Neural Computation*, *22*, 2655–2677.

Zhang, H., Liu, X., Xu, D., & Zhang, Y. (2014). Convergence analysis of fully complex backpropagation algorithm based on Wirtinger calculus. *Cognitive Neurodynamics*, *8*, 261–266.

Zhang, H., Xu, D., & Zhang, Y. (2014). Boundedness and convergence of split-complex back-propagation algorithm with momentum and penalty. *Neural Processing Letters*, *39*, 297–307.

Zhang, H., Zhang, C., & Wu, W. (2009). Convergence of batch split-complex backpropagation algorithm for complex-valued neural networks. *Discrete Dynamics in Nature and Society*, *2009*, 1–16.