BRIEF COMMUNICATION

Convergence analysis of fully complex backpropagation algorithm based on Wirtinger calculus

Huisheng Zhang · Xiaodong Liu · Dongpo Xu · Ying Zhang

Abstract This paper considers the fully complex backpropagation algorithm (FCBPA) for training the fully complex-valued neural networks. We prove both the weak convergence and strong convergence of FCBPA under mild conditions. The decreasing monotonicity of the error functions during the training process is also obtained. The derivation and analysis of the algorithm are under the framework of Wirtinger calculus, which greatly reduces the description complexity. The theoretical results are substantiated by a simulation example.

Keywords Complex-valued neural networks · Fully complex backpropagation algorithm · Wirtinger calculus · Convergence

Introduction

The theoretical studies and practical implementations of complex-valued neural network (CVNN) have attracted considerable attention in signal processing, pattern recognition, and medical information processing (Fink et al. 2014; Hirose 2012; Nitta 2013). Based on different choices

H. Zhang (✉) · Y. Zhang
Department of Mathematics, Dalian Maritime University,
Dalian 116026, People's Republic of China
e-mail: zhhuisheng@163.com

H. Zhang · X. Liu
Research Center of Information and Control, Dalian University
of Technology, Dalian 116024, People's Republic of China

D. Xu
College of Science, Harbin Engineering University,
Harbin 150001, People's Republic of China

of the activation function, there are two main CVNN models: the split CVNN (Nitta 1997) and the fully CVNN (Kim and Adali 2003). The split CVNN uses a pair of real-valued functions to separately process the real part and the imaginary part of the neuron's input signal. This strategy can effectively overcome the singularity problem during the training procedure. In contrast, activation functions of the fully CVNN are fully complex-valued, which helps the network make fully use of the phase information and thus achieve better performance in some applications. As one of the most popular training methods for neural networks, backpropagation algorithm (BPA) has been extended from the real domain to the complex domain in order to train the CVNN. Accordingly, there are two types of complex BPA: one is the split-complex BPA (SCBPA) (Nitta 1997) for split CVNN, and another is the fully complex BPA (FCBPA) (Li and Adali 2008) for fully CVNN.

Convergence is the precondition for any iteration algorithm to be used in real applications. (Wei et al. 2013; Osborn 2010) The convergence of the BPA has been extensively studied in literature (Wu et al. 2005, 2011; Zhang et al. 2007, 2008, 2009; Shao and Zheng 2011), where the boundedness and differentiability of the activation function are usually two necessary conditions for the convergence analysis. However, as stated by the Liouvilles theorem (an entire and bounded function in the complex domain is a constant), the complex activation function can not be both bounded and analytic. The conflict between the boundedness and differentiability of the activation function makes the theoretical convergence analysis for the complex BPA more difficult than that for the BPA. Fortunately, as the activation functions of the split CVNN can be split into two bounded and differential real-valued functions, the convergence analysis of SCBPA can then be conducted in the real domain. For the corresponding convergence results,

we refer to (Nitta 1997; Zhang et al. 2013; Xu et al. 2010). However, though FCBPA has been widely used in many applications and has been experimentally shown to be convergent for some kinds of activation functions (Kim and Adali 2003), the theoretical convergence analysis remains challengeable.

Despite the challenge posed by the Liouville's theorem, another challenge for the theoretical convergence analysis of FCBPA is that the traditional mean value theorem, which is vital for the convergence analysis of BPA, dos not hold in the complex domain. (For example: $f(z) = e^z$ $with z_2 = z_1 + 2\pi i$. We have $f(z_2) - f(z_1) = 0$ but $(z_2 - z_1)f'(w) = 2\pi i e^w \neq 0$ for all w.) By expanding the analytic function with Taylor series and omitting the high order terms, some local stability results for complex ICA are obtained (Adali et al. 2008). Under the assumption that the activation function is a contraction, the convergence of the complex nonlinear adaptive filters is proved (Mandic and Goh 2009). However, to the best of our knowledge, the theoretical convergence results of the FCBPA has not yet been established. This becomes the main concern of this paper. Specifically, we make the following contributions:

- By introducing a mean value theorem for a holomorphic function (Mcleod 1965), we will prove both the weak convergence and the strong convergence of FCBPA.
- Instead of dropping the high order terms of the Taylor series (Adali et al. 2008), we give an accurate estimation for the difference of the error function between two iterations using the mean value theorem. As a result, our results are of global nature in that they are valid for arbitrarily given initial values of the weights.
- The restrictive condition that the activation function is a contraction is not needed in our analysis.
- The derivation and analysis of the algorithm are under the framework of Wirtinger calculus, which greatly reduces the description complexity.

The remainder of this paper is organized as follows. The network structure and the derivation of the FCBPA based on Wirtinger calculus are described in the next section. "Main results" section presents the main convergence theorem of the paper. The detailed proof of the theorem is given in Section Proofs. In "Simulation result" section we use a simulation example to support our theoretical results. The paper is concluded in "Conclusion" section.

Network structure and FCBPA based on Wirtinger calculus

We consider a single hidden layer feedforward network consisting of p input nodes, q hidden nodes, and 1 output

node. Let $\mathbf{w}_0 = (w_{01}, w_{02}, \cdots, w_{0q})^T \in \mathbb{C}^q$ be the weight vector between all the hidden units and the output unit, and $\mathbf{w}_l = (w_{l1}, w_{l2}, \cdots, w_{lp})^T \in \mathbb{C}^p$ be the weight vector between all the input units and the hidden unit l ($l = 1, 2, \cdots, q$). To simplify the presentation, we write all the weight parameters in a compact form, i.e., $\mathbf{w} = (\mathbf{w}_0^T, \mathbf{w}_1^T, \cdots, \mathbf{w}_q^T)^T \in \mathbb{C}^{q+pq}$ and we define a matrix $\mathbf{V} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_q)^T \in \mathbb{C}^{q \times p}$.

Given activation functions $f, g : \mathbb{C} \to \mathbb{C}$ for the hidden layer and output layer, respectively, we define a vector function $\mathbf{F}(\mathbf{x}) = (f(x_1), f(x_2), \cdots, f(x_q))^T$ for $\mathbf{x} = (x_1, x_2, \cdots, x_q)^T \in \mathbb{C}^q$. For an input $z \in \mathbb{C}^p$, the output vector of the hidden layer can be written as $\mathbf{F}(\mathbf{V}z)$ and the final output of the network can be written as

$$y = g(\mathbf{w}_0 \cdot \mathbf{F}(\mathbf{V}z)), \tag{1}$$

where $\mathbf{w}_0 \cdot \mathbf{F}(\mathbf{V}z)$ represents the inner product between the two vectors $\mathbf{w}_0$ and $\mathbf{F}(\mathbf{V}z)$.

Suppose that $\{z^k, d^k\}_{k=1}^K \subset \mathbb{C}^p \times \mathbb{C}$ is a given set of training samples, where z^k is the input, and d^k is the desired output. The aim of the network training is to find the appropriate network weights $\mathbf{w}^\star$ that can minimize the error function

$$E(\mathbf{w}) = \sum_{k=1}^K (d_k - y_k)(\overline{d_k} - \overline{y_k}), \tag{2}$$

where

$$y_k = g(\mathbf{w}_0 \cdot \mathbf{F}(\mathbf{V}z^k)) \tag{3}$$

and the notation $^-$ denotes the complex conjugation.

As noted by Adali et al. (2008), any function $h(z)$ that is analytic in a bounded zone $|z| < R$ with a Taylor series expansion with all real coefficients in $|z| < R$ satisfies the property $\overline{h(z)} = h(\overline{z})$. Examples of such functions include polynomials and most trigonometric functions and their hyperbolic counterparts, which are qualified activation functions of CVNNs (Kim and Adali 2003). As a result, we suppose both the activation functions $f(\cdot)$ and $g(\cdot)$ satisfy $\overline{f(z)} = f(\overline{z})$, $\overline{g(z)} = g(\overline{z})$. Therefore,

$$\overline{y_k} = g(\overline{\mathbf{w}_0} \cdot \mathbf{F}(\overline{\mathbf{V}z^k})). \tag{4}$$

$E(\mathbf{w})$ can be viewed as a function of complex variable vector $\mathbf{w}$ and its conjugate $\overline{\mathbf{w}}$. According to the Wirtinger calculus (Brandwood 1983; Bos 1994), we can define two gradient vectors $\nabla_{\mathbf{w}}E$ (by taking partial derivatives with respect to $\mathbf{w}$ at the same time treating $\overline{\mathbf{w}}$ as a constant vector in E) and $\nabla_{\overline{\mathbf{w}}}E$ (by taking partial derivatives with respect to $\overline{\mathbf{w}}$ at the same time treating $\mathbf{w}$ as a constant vector). Then the gradient $\nabla_{\overline{\mathbf{w}}}E$ defines the direction of the maximum rate of change in $E(\mathbf{w})$ with respect to $\mathbf{w}$. As the output node y_k does not explicitly contain the variable vector $\overline{\mathbf{w}}$, we can conclude $\nabla_{\overline{\mathbf{w}}}y_k = 0$. Thus, by the chain rule of the Wirtinger calculus, we have

$$\frac{\partial E(\mathbf{w})}{\partial \overline{\mathbf{w}_0}} = \sum_{k=1}^{K}(y_k - d_k)g'(\overline{\mathbf{w}_0} \cdot \mathbf{F}(\overline{\mathbf{V}\mathbf{z}^k}))\mathbf{F}(\overline{\mathbf{V}\mathbf{z}^k}), \tag{5}$$

$$\frac{\partial E(\mathbf{w})}{\partial \overline{\mathbf{w}_l}} = \sum_{k=1}^{K}(y_k - d_k)g'(\overline{\mathbf{w}_0} \cdot \mathbf{F}(\overline{\mathbf{V}\mathbf{z}^k}))\overline{w_{0l}}f'(\overline{\mathbf{w}_l} \cdot \overline{\mathbf{z}^k})\overline{\mathbf{z}^k},$$
$$l = 1, 2 \cdots, q. \tag{6}$$

Obviously,

$$\nabla_{\overline{\mathbf{w}}}E(\mathbf{w}) = \left(\left(\frac{\partial E(\mathbf{w})}{\partial \overline{\mathbf{w}_0}}\right)^T, \left(\frac{\partial E(\mathbf{w})}{\partial \overline{\mathbf{w}_1}}\right)^T, \cdots, \left(\frac{\partial E(\mathbf{w})}{\partial \overline{\mathbf{w}_q}}\right)^T \right)^T. \tag{7}$$

Starting from an arbitrary initial value $\mathbf{w}^0$, the BPA based on Wirtinger calculus updates the weights $\{\mathbf{w}^n\}$ iteratively by

$$\mathbf{w}^{n+1} = \mathbf{w}^n - \eta\nabla_{\overline{\mathbf{w}}}E(\mathbf{w}^n), \tag{8}$$

where $\eta > 0$ is the learning rate.

Main results

The following assumptions are needed in our convergence analysis.

(A1) There exists a constant $c_1 > 0$ such that $\|\mathbf{w}_l^n\| \le c_1$ for all $l = 0, 1, \cdots, q$, $n = 0, 1, \cdots$;

(A2) The functions $f(z)$ and $g(z)$ are analytic in a bounded zone $|z| < R$ with a Taylor series expansion with all real coefficients in $|z| < R$, where $R > \max\{c_2, c_3\}$ (c_2 and c_3 are defined in (17) below).

(A3) The set $\Phi_1 = \{\mathbf{w} : \nabla_{\overline{\mathbf{w}}}E(\mathbf{w}) = 0\}$ contains only finite points.

Remark 1 Assumption (A1) is the usual condition for the convergence analysis of the gradient method for both the real-valued neural networks (Zhang et al. 2007, 2008) and the CVNNs (Xu et al. 2010) in literature. As noted by Adali et al. (2008), Assumption (A2) is satisfied by quite a number of functions which are qualified as activation functions of the fully CVNNs. Assumption (A3) is used to establish a strong convergence result.

Now we present our convergence results.

Theorem 1 *Suppose that the error function is given by (2), that the weight sequence $\{\mathbf{w}^n\}$ is generated by the algorithm (8) for any initial value $\mathbf{w}^0$, that $0 < \eta < \frac{1}{L}$, where L is defined by (27) below, and that Assumptions (A1) and (A2) are valid. Then we have*

(a)

$$E(\mathbf{w}^{n+1}) \le E(\mathbf{w}^n), \quad n = 0, 1, 2, \cdots; \tag{9}$$

(b)

There is $E^{\star} > 0$ such that $\lim_{n \to \infty} E(\mathbf{w}^n) = E^{\star}$; $\tag{10}$

(c)

There holds the weak convergence: $\lim_{n \to \infty} \|\nabla_{\overline{w}}E(w^n)\| = 0.$
$$\tag{11}$$

Moreover, if Assumption (A3) is valid, then there holds the strong convergence: there exists a point $\mathbf{w}^{\star} \in \Phi$ such that
(d)

$$\lim_{n \to \infty} \mathbf{w}^n = \mathbf{w}^{\star}. \tag{12}$$

Proofs

Lemma 1 *[see Theorem 10 in by Mcleod (1965)] Suppose h is a holomorphic function defined on a connected open set G in the complex plane. If z_1 and z_2 are points in G such that the segment joining them is also in G then*

$$h(z_2) - h(z_1) = (z_2 - z_1)(\lambda_1 h'(\xi_1) + \lambda_2 h'(\xi_2)) \tag{13}$$

for some ξ_1 and ξ_2 on the segment joining z_1 and z_2 and some nonnegative numbers λ_1 and λ_2 such that $\lambda_1 + \lambda_2 = 1$.

Lemma 2 *Suppose Assumptions (A1) and (A2) are valid, then $\nabla_{\overline{\mathbf{w}}}E(\mathbf{w})$ satisfies Lipschitz condition, that is, there exists a positive constant L_1, such that*

$$\|\nabla_{\overline{\mathbf{w}}}E(\mathbf{w}^{n+1}) - \nabla_{\overline{\mathbf{w}}}E(\mathbf{w}^n)\| \le L_1\|\mathbf{w}^{n+1} - \mathbf{w}^n\|. \tag{14}$$

Similarly, there exists a positive constant L_2, such that

$$\|\nabla_{\mathbf{w}}E(\mathbf{w}^{n+1}) - \nabla_{\mathbf{w}}E(\mathbf{w}^n)\| \le L_2\|\mathbf{w}^{n+1} - \mathbf{w}^n\|. \tag{15}$$

Proof For simplicity, we introduce the following notations:

$$\mathbf{F}^{n,k} = \mathbf{F}(\mathbf{V}^n\mathbf{z}^k), \overline{\mathbf{F}}^{n,k} = \mathbf{F}(\overline{\mathbf{V}^n\mathbf{z}^k}), \tag{16}$$

for $n = 1, 2, \cdots, k = 1, 2, \cdots, K$.

By Assumption (A2), f and g have differentials of any order in the zone $\{z: |z| < R\}$. In addition, recalling $\{z^k, d^k\}_{k=1}^{K}$ is finite and $\{\mathbf{w}^n\}$ is bounded, we can define c_2 and c_3 such that

$$c_2 = \sup_{l,n,k}|\overline{\mathbf{w}_l^n} \cdot \overline{\mathbf{z}^k}|, c_3 = \sup_{n,k}|\overline{\mathbf{w}_0^n} \cdot \mathbf{F}(\overline{\mathbf{V}^n\mathbf{z}^k})|, \tag{17}$$

$$c_4 = \sup_{|z| < \max\{c_2,c_3\}} \{|f(z)|, |g(z)|, |f'(z)|, |g'(z)|, |g''(z)|\}. \tag{18}$$

Using (18), Lemma 1 and the Cauchy-Schwartz Inequality, for any $1 \le k \le K$ and $n = 0, 1, \cdots$, we have that

$$\|\mathbf{F}^{n+1,k} - \mathbf{F}^{n,k}\| = \left\| \begin{pmatrix} f(\mathbf{w}_1^{n+1} \cdot \mathbf{z}^k) - f(\mathbf{w}_1^n \cdot \mathbf{z}^k) \\ \vdots \\ f(\mathbf{w}_q^{n+1} \cdot \mathbf{z}^k) - f(\mathbf{w}_q^n \cdot \mathbf{z}^k) \end{pmatrix} \right\|$$

$$= \left\| \begin{pmatrix} (\mathbf{w}_1^{n+1} \cdot \mathbf{z}^k - \mathbf{w}_1^n \cdot \mathbf{z}^k)(\lambda_{11}f'(\xi_{11}) + \lambda_{12}f'(\xi_{12})) \\ \vdots \\ (\mathbf{w}_q^{n+1} \cdot \mathbf{z}^k - \mathbf{w}_q^n \cdot \mathbf{z}^k)(\lambda_{q1}f'(\xi_{q1}) + \lambda_{q2}f'(\xi_{q2})) \end{pmatrix} \right\|$$

$$\leq c_5 \left\| \begin{pmatrix} \mathbf{w}_1^{n+1} - \mathbf{w}_1^n \\ \vdots \\ \mathbf{w}_q^{n+1} - \mathbf{w}_q^n \end{pmatrix} \right\| \leq c_5 \sum_{l=1}^q \|\mathbf{w}_l^{n+1} - \mathbf{w}_l^n\|, \quad (19)$$

where $c_5 = \sqrt{q}c_4 \sup \|\mathbf{z}^k\|$, $\lambda_{l1} \geq 0$, $\lambda_{l2} \geq 0$, $\lambda_{l1} + \lambda_{l2} = 1$, ξ_{l1} and ξ_{l2} lie on the segment joining $\mathbf{w}_l^{n+1} \cdot \mathbf{z}^k$ and $\mathbf{w}_l^n \cdot \mathbf{z}^k$, $l = 1, \cdots, q$.

Similarly, we have

$$\|\overline{\mathbf{F}}^{n+1,k} - \overline{\mathbf{F}}^{n,k}\| \leq c_5 \sum_{l=1}^q \|\mathbf{w}_l^{n+1} - \mathbf{w}_l^n\|. \quad (20)$$

By (18), (19), Lemma 1 and the Cauchy-Schwartz Inequality we have that for any $1 \leq k \leq K$ and $n = 0, 1, \cdots$

$$\|y_k^{n+1} - y_k^n\| = \|g(\mathbf{w}_0^{n+1} \cdot \mathbf{F}^{n+1,k}) - g(\mathbf{w}_0^n \cdot \mathbf{F}^{n,k})\|$$

$$= \|(\mathbf{w}_0^{n+1} \cdot \mathbf{F}^{n+1,k} - \mathbf{w}_0^n \cdot \mathbf{F}^{n,k})(\eta_1 g'(\zeta_1) + \eta_2 g'(\zeta_2))\|$$

$$\leq c_4(\|\mathbf{w}_0^{n+1} - \mathbf{w}_0^n\|\|\mathbf{F}^{n+1,k}\| + \|\mathbf{w}_0^n\|\|\mathbf{F}^{n+1,k} - \mathbf{F}^{n,k}\|)$$

$$\leq \sqrt{q}c_4^2\|\mathbf{w}_0^{n+1} - \mathbf{w}_0^n\| + c_1 c_4 c_5 \sum_{l=1}^q \|\mathbf{w}_l^{n+1} - \mathbf{w}_l^n\|$$

$$\leq c_6\|\mathbf{w}^{n+1} - \mathbf{w}^n\|, \quad (21)$$

where $c_6 = \sqrt{q+1} \max\{\sqrt{q}c_4^2, c_1 c_4 c_5\}$, $\eta_1 \geq 0, \eta_2 \geq 0, \eta_1 + \eta_2 = 1$, ζ_1 and ζ_2 lie on the segment joining $\mathbf{w}_0^{n+1} \cdot \mathbf{F}^{n+1,k}$ and $\mathbf{w}_0^n \cdot \mathbf{F}^{n,k}$.

In the same way, we can prove that

$$\|g'(\overline{\mathbf{w}_0^{n+1}} \cdot \overline{\mathbf{F}}^{n+1,k}) - g'(\overline{\mathbf{w}_0^n} \cdot \overline{\mathbf{F}}^{n,k})\| \leq c_6\|\mathbf{w}^{n+1} - \mathbf{w}^n\|. \quad (22)$$

With (18), (20), (22), and Cauchy-Schwartz Inequality we obtain

$$\|g'(\overline{\mathbf{w}_0^{n+1}} \cdot \overline{\mathbf{F}}^{n+1,k})\overline{\mathbf{F}}^{n+1,k} - g'(\overline{\mathbf{w}_0^n} \cdot \overline{\mathbf{F}}^{n,k})\overline{\mathbf{F}}^{n,k}\|$$

$$= \|(g'(\overline{\mathbf{w}_0^{n+1}} \cdot \overline{\mathbf{F}}^{n+1,k}) - g'(\overline{\mathbf{w}_0^n} \cdot \overline{\mathbf{F}}^{n,k}))\overline{\mathbf{F}}^{n+1,k}$$

$$+ g'(\overline{\mathbf{w}_0^n} \cdot \overline{\mathbf{F}}^{n,k})(\overline{\mathbf{F}}^{n+1,k} - \overline{\mathbf{F}}^{n,k})\|$$

$$\leq |g'(\overline{\mathbf{w}_0^{n+1}} \cdot \overline{\mathbf{F}}^{n+1,k}) - g'(\overline{\mathbf{w}_0^n} \cdot \overline{\mathbf{F}}^{n,k})|\|\overline{\mathbf{F}}^{n+1,k}\| \quad (23)$$

$$+ |g'(\overline{\mathbf{w}_0^n} \cdot \overline{\mathbf{F}}^{n,k})|\|(\overline{\mathbf{F}}^{n+1,k} - \overline{\mathbf{F}}^{n,k})\|$$

$$\leq c_6\sqrt{q}c_4\|\mathbf{w}^{n+1} - \mathbf{w}^n\| + c_4 c_5 \sum_{l=1}^q \|\mathbf{w}_l^{n+1} - \mathbf{w}_l^n\|$$

$$\leq \sqrt{q}c_4(c_5 + c_6)\|\mathbf{w}^{n+1} - \mathbf{w}^n\|$$

Combining (18), (21), (23), and the Cauchy-Schwartz Inequality we can conclude

$$\|\frac{\partial E(\mathbf{w}^{n+1})}{\partial \overline{\mathbf{w}_0}} - \frac{\partial E(\mathbf{w}^n)}{\partial \overline{\mathbf{w}_0}}\|$$

$$= \| \sum_{k=1}^K (y_k^{n+1} - d_k)g'(\overline{\mathbf{w}_0^{n+1}} \cdot \overline{\mathbf{F}}^{n+1,k})\overline{\mathbf{F}}^{n+1,k}$$

$$- \sum_{k=1}^K (y_k^n - d_k)g'(\overline{\mathbf{w}_0^n} \cdot \overline{\mathbf{F}}^{n,k})\overline{\mathbf{F}}^{n,k}\|$$

$$= \| \sum_{k=1}^K ((y_k^{n+1} - y_k^n)g'(\overline{\mathbf{w}_0^{n+1}} \cdot \overline{\mathbf{F}}^{n+1,k})\overline{\mathbf{F}}^{n+1,k}$$

$$+ (y_k^n - d_k)(g'(\overline{\mathbf{w}_0^{n+1}} \cdot \overline{\mathbf{F}}^{n+1,k})\overline{\mathbf{F}}^{n+1,k} - g'(\overline{\mathbf{w}_0^n} \cdot \overline{\mathbf{F}}^{n,k})\overline{\mathbf{F}}^{n,k}))\|$$

$$\leq \sum_{k=1}^K (|(y_k^{n+1} - y_k^n)g'(\overline{\mathbf{w}_0^{n+1}} \cdot \overline{\mathbf{F}}^{n+1,k})|\|\overline{\mathbf{F}}^{n+1,k}\|$$

$$+ |y_k^n - d_k|\|g'(\overline{\mathbf{w}_0^{n+1}} \cdot \overline{\mathbf{F}}^{n+1,k})\overline{\mathbf{F}}^{n+1,k} - g'(\overline{\mathbf{w}_0^n} \cdot \overline{\mathbf{F}}^{n,k})\overline{\mathbf{F}}^{n,k}\|)$$

$$\leq \sum_{k=1}^K (c_4^2\sqrt{q}|y_k^{n+1} - y_k^n| + (c_4 + \sup d_k)\|g'(\overline{\mathbf{w}_0^{n+1}} \cdot \overline{\mathbf{F}}^{n+1,k})$$

$$\overline{\mathbf{F}}^{n+1,k} - g'(\overline{\mathbf{w}_0^n} \cdot \overline{\mathbf{F}}^{n,k})\overline{\mathbf{F}}^{n,k}\|) \leq L_3\|\mathbf{w}^{n+1} - \mathbf{w}^n\| \quad (24)$$

where $L_3 = K\sqrt{q}c_4(c_4 c_6 + (c_4 + \sup d_k)(c_5 + c_6))$.

Similarly, there exists a Lipschitz constants L_4 such that for $l = 1, \cdots, q$

$$\|\frac{\partial E(\mathbf{w}^{n+1})}{\partial \overline{\mathbf{w}_l}} - \frac{\partial E(\mathbf{w}^n)}{\partial \overline{\mathbf{w}_l}}\| \leq L_4\|\mathbf{w}^{n+1} - \mathbf{w}^n\|. \quad (25)$$

Hence, (7), (24), and (25) validate (14) by setting $L_1 = L_3 + qL_4$.

Equation (15) can be proved in a similar way to (14). □

Now, we proceed to the proof of Theorem 2 by dealing with Equations (9)-(12) separately.

Proof of (9) By the differential mean value theorem, there exists a constant $\theta \in [0, 1]$, such that

$$E(\mathbf{w}^{n+1}) - E(\mathbf{w}^n)$$

$$= (\nabla_{\mathbf{w}}E(\mathbf{w}^n + \theta\Delta\mathbf{w}^n))^T\Delta\mathbf{w}^n + (\nabla_{\overline{\mathbf{w}}}E(\mathbf{w}^n + \theta\Delta\mathbf{w}^n))^T\overline{\Delta\mathbf{w}^n}$$

$$= (\nabla_{\mathbf{w}}E(\mathbf{w}^n))^T\Delta\mathbf{w}^n + (\nabla_{\overline{\mathbf{w}}}E(\mathbf{w}^n))^T\overline{\Delta\mathbf{w}^n}$$

$$+ (\nabla_{\mathbf{w}}E(\mathbf{w}^n + \theta\Delta\mathbf{w}^n) - \nabla_{\mathbf{w}}E(\mathbf{w}^n))^T\Delta\mathbf{w}^n$$

$$+ (\nabla_{\overline{\mathbf{w}}}E(\mathbf{w}^n + \theta\Delta\mathbf{w}^n) - \nabla_{\overline{\mathbf{w}}}E(\mathbf{w}^n))^T\overline{\Delta\mathbf{w}^n}$$

$$\leq 2Re(\nabla_{\mathbf{w}}E(\mathbf{w}^n)^T\Delta\mathbf{w}^n)$$

$$+ (\|\nabla_{\mathbf{w}}E(\mathbf{w}^n + \theta\Delta\mathbf{w}^n) - \nabla_{\mathbf{w}}E(\mathbf{w}^n)\|$$

$$+ \|\nabla_{\overline{\mathbf{w}}}E(\mathbf{w}^n + \theta\Delta\mathbf{w}^n) - \nabla_{\overline{\mathbf{w}}}E(\mathbf{w}^n)\|)\|\Delta\mathbf{w}^n\|$$

$$\leq -2\eta Re((\nabla_{\overline{\mathbf{w}}}E(\mathbf{w}^n))^H\nabla_{\overline{\mathbf{w}}}E(\mathbf{w}^n))$$

$$+ (L_1 + L_2)\theta\|\Delta\mathbf{w}^n\|^2$$

$$= (-2\eta + (L_1 + L_2)\theta\eta^2)\|\nabla_{\overline{\mathbf{w}}}E(\mathbf{w}^n)\|^2 \quad (26)$$

To make (9) valid, we only require the learning rate η to satisfy

$$0 < \eta < L, \tag{27}$$

where $L = \frac{2}{(L_1 + L_2)\theta}$.   □

Proof of (10)   Equation (10) is directly obtained by (9) and $E(\mathbf{w}^n) > 0 (n = 1, 2, \cdots)$.

Proof of (11)   Let $\beta = 2\eta - (L_1 + L_2)\theta\eta^2$. By (26), we have

$$\begin{aligned} E(\mathbf{w}^{n+1}) &\leq E(\mathbf{w}^n) - \beta \|\nabla_{\overline{\mathbf{w}}} E(\mathbf{w}^n)\|^2 \\ &\leq \cdots \leq E(\mathbf{w}^0) - \beta \sum_{t=0}^{n} \|\nabla_{\overline{\mathbf{w}}} E(\mathbf{w}^t)\|^2. \end{aligned} \tag{28}$$

Considering $E(\mathbf{w}^{n+1}) > 0$, let $n \to \infty$, then we have

$$\beta \sum_{t=0}^{\infty} \|\nabla_{\overline{\mathbf{w}}} E(\mathbf{w}^t)\|^2 \leq E(\mathbf{w}^0) < \infty. \tag{29}$$

This immediately gives

$$\lim_{n \to \infty} \|\nabla_{\overline{\mathbf{w}}} E(\mathbf{w}^n)\| = 0. \tag{30}$$

□

The following lemma, which will be used in the proof of (12), is a generalization of Theorem 14.1.5 by Ortega and Rheinboldt (1970) from the real domain to the complex domain. The proof of this lemma follows the same route as (Ortega and Rheinboldt 1970) and we omit it here.

Lemma 3 [22] *Let* $\varphi : \Phi \subset \mathbb{C}^k \to \mathbb{C}(k \geq 1)$ *be continuous for a bounded closed region* Φ, *and* $\Phi_0 = \{\mathbf{z} \in \Phi : \varphi(\mathbf{z}) = 0\}$. *Suppose the set* Φ_0 *contains only finite points and the sequence* $\{\mathbf{z}^n\}$ *satisfy*:

(i) $\lim_{n \to \infty} \varphi(\mathbf{z}^n) = 0$;
(ii) $\lim_{n \to \infty} \|\mathbf{z}^{n+1} - \mathbf{z}^n\| = 0$.

Then, there exists a unique $\mathbf{z}^{\star} \in \Phi_0$ *such that* $\lim_{n \to \infty} \mathbf{z}^n = \mathbf{z}^{\star}$.

Proof of (12)   Obviously $\nabla_{\overline{\mathbf{w}}} E(\mathbf{w})$ is continuous under the Assumption (A2). Using (8) and (11), we have

$$\lim_{n \to \infty} \|\mathbf{w}^{n+1} - \mathbf{w}^n\| = \eta \lim_{n \to \infty} \|\nabla_{\overline{\mathbf{w}}} E(\mathbf{w}^n)\| = 0. \tag{31}$$

Furthermore, the Assumption (A3) is valid. Thus, applying Lemma 3, there exists a unique $\mathbf{w}^{\star} \in \Phi$ such that $\lim_{n \to \infty} \mathbf{w}^n = \mathbf{w}^{\star}$.   □

Simulation result

In this section we illustrate the convergence behavior of the FCBPA by the problem of one-step-ahead prediction of the complex-valued nonlinear signals. The nonlinear benchmark input signal is given by (Mandic and Goh 2009)
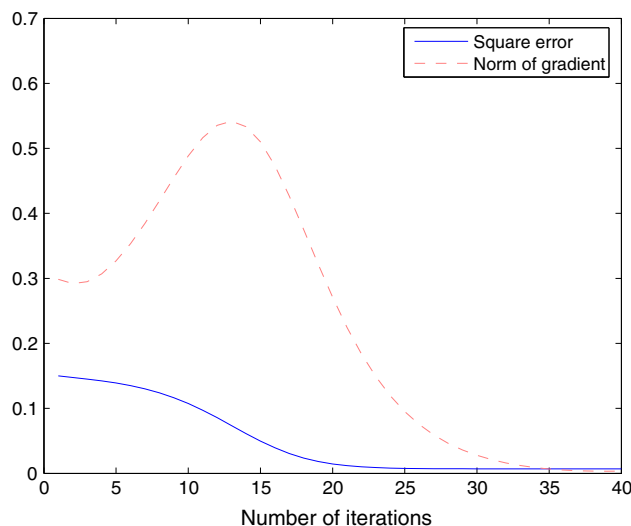
Fig. 1 Convergence behavior of FCBPA

$$z(t) = \frac{z(t-1)}{1 + z^2(t-1)} + n^3(t), \tag{32}$$

where $n(t)$ is a complex white Gaussian noise with zero mean and unit variance.

This example uses a network with one input node, five hidden nodes, and one output node. We set the activation function for both the hidden layer and output layer to be $sin(\cdot)$, which is analytic in the complex domain. The learning rate η is set to be 0.1. The test is carried out with the initial weights (both the real part and the imaginary part) taken as random numbers from the interval $[-0.1, 0.1]$. The simulation results are presented in Fig. 1, which shows that the gradient tends to zero and the square error decreases monotonically as the number of iteration increases and at last it tends to a constant. This supports our theoretical findings.

Conclusion

In this paper, under the framework of Wirtinger calculus, we investigate the FCBPA for fully CVNN. Using a mean value theorem for holomorphic functions, under mild conditions we prove the gradient of the error function with respect to the network weight vector satisfies the Lipschitz condition. Based on this conclusion, both the weak convergence and strong convergence of the algorithm are proved. Simulation results substantiate the theoretical findings.

References

Adali T, Li H, Novey M et al (2008) Complex ICA using nonlinear functions. IEEE Trans Signal Process 56(9):4356–544

Bos AVD (1994) Complex gradient and Hessian. Proc Inst Electr Eng Vision Image Signal Process 141:380–382

Brandwood D (1983) Complex gradient operator and its application in adaptive array theory. Proc Inst Electr Eng 130:11–16

Fink O, Zio E, Weidmann U (2014) Predicting component reliability and level of degradation with complex-valued neural networks. Reliab Eng Syst Safe 121:198–206

Hirose A (2012) Complex-valued neural networks. Springer-Verlag, Berlin Heidelberg

Kim T, Adali T (2003) Approximation by fully complex multilayer perceptrons. Neural Comput 15:1641–666

Li H, Adali T (2008) Complex-valued adaptive signal processing using nonlinear functions. EURASIP J Adv Signal Process 2008:122

Mandi DP, Goh SL (2009) Complex valued nonlinear adaptive filters. Wiley, Chichester

Mcleod RM (1965) Mean value theorems for vector valued functions. Proc Edinburgh Math Soc 14(2):197–209

Nitta T (1997) An extension of the back-propagation algorithm to complex numbers. Neural Netw 10(8): 1391–1415

Nitta T (2013) Local minima in hierarchical structures of complex-valued neural networks. Neural Netw 43: 1–7

Osborn GW (2010) A Kalman filtering approach to the representation of kinematic quantities by the hippocampal-entorhinal complex. Cogn Neurodyn 4:C315–C335

Shao HM, Zheng GF (2011) Boundedness and convergence of online gradient method with penalty and momentum. Neurocomputing 74:765–770

Wei H, Ren Y, Wang ZY (2013) A computational neural model of orientation detection based on multiple guesses: comparison of geometrical and algebraic models. Cogn Neurodyn 7:C361–C379

Wu W, Feng GR, Li ZX, et al. (2005) Deterministic convergence of an online gradient method for BP neural networks. IEEE Trans Neural Netw 16:533–540

Wang J, Wu W, Zurada J (2011) Deterministic convergence of conjugate gradient method for feedforward neural networks. Neurocomputing 74:2368–2376

Xu DP, Zhang HS, Liu L (2010) Convergence analysis of three classes of split-complex gradient algorithms for complex-valued recurrent neural networks. Neural Comput 22(10):2655-2677

Zhang C, Wu W, Xiong Y (2007) Convergence analysis of batch gradient algorithm for three classes of sigma-pi neural networks. Neural Process Lett 26:177–180

Zhang C, Wu W, Chen XH, et al. (2008) Convergence of BP algorithm for product unit neural networks with exponential weights. Neurocomputing 72:513–520

Zhang HS, Wu W, Liu F, Yao MC (2009) Boundedness and convergence of online gradient method with penalty for feedforward neural networks. IEEE Trans Neural Netw 20(6):1050–1054

Zhang HS, Xu DP, Zhang Y (2013) Boundedness and convergence of split-complex back-propagation algorithm with momentum and penalty. Neural Process Lett. doi:10.1007/s11063-013-9305-x

Ortega JM, Rheinboldt WC (1970) Iterative solution of nonlinear equations in several variables. Academic Press, New York