Cogn Neurodyn (2015) 9:331–340
DOI 10.1007/s11571-014-9323-z

RESEARCH ARTICLE

Deterministic convergence of chaos injection-based gradient method for training feedforward neural networks

Huisheng Zhang · Ying Zhang · Dongpo Xu · Xiaodong Liu

Received: 13 October 2014 / Revised: 26 November 2014 / Accepted: 10 December 2014 / Published online: 1 January 2015

Abstract It has been shown that, by adding a chaotic sequence to the weight update during the training of neural networks, the chaos injection-based gradient method (CIBGM) is superior to the standard backpropagation algorithm. This paper presents the theoretical convergence analysis of CIBGM for training feedforward neural networks. We consider both the case of batch learning as well as the case of online learning. Under mild conditions, we prove the weak convergence, i.e., the training error tends to a constant and the gradient of the error function tends to zero. Moreover, the strong convergence of CIBGM is also obtained with the help of an extra condition. The theoretical results are substantiated by a simulation example.

Keywords Feedforward neural networks · Chaos injection-based gradient method · Batch learning · Online learning · Convergence

Introduction

Gradient method (GM) has been widely used as a training algorithm for feedforward neural networks. GM can be implemented in two practical ways: the batch learning and the online learning (Haykin 2008). The batch learning approach accumulates the weight correction over all the training samples before actually performing the update, nevertheless the online learning approach updates the network weights immediately after each training sample is fed. Though GM is widely used in neural network fields, it also has drawbacks of slow learning and getting trapped in local minimum. To overcome those problems, many heuristic improvements have been proposed, such as adding a penalty term to the error function (Karnin 1990), adding a momentum to the weight update (Zhang et al. 2006), injecting noise into the learning procedure (Sum et al. 2012a, b; Ho et al. 2010), etc. Some other nonlinear optimization algorithms such as the Newton method (Osowski et al. 1996), conjugate-gradient method (Charalambous 1992), extended Kalman filtering (Iiguni et al. 1992), and Levenberg–Marquardt method (Hagan and Mehnaj 1994) have also been used for training neural networks. Though these algorithms converge in fewer iterations than GM, they require much more computation per pattern, which makes them not so suitable especially for online learning (Behera et al. 2006). Thus, gradient method remains attractive because of its simplicity and ease of implementation.

As convergence is a precondition for the practical usage of a learning algorithm, the convergence analysis of GM and its various modifications have attracted many researchers in neural network fields (Fine and Mukherjee 1999; Wu et al. 2005, 2011; Wang et al. 2011; Shao and Zheng 2011; Zhang et al. 2007, 2008, 2009, 2012, 2014; Fan et al. 2014; Yu and Chen 2012). Recently, Sum John, Leung Chi-Sing and Ho Kevin theoretically investigated the convergence of noise injection-based online gradient methods (NIBOGM) in Sum et al. (2012a, b) and Ho et al. (2010), where the noises are independent mean zero

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

Gaussian-distributed random variables. For the stability of the noise-induced neural systems and the effects of the noise on neural networks, we refer to Wu et al. (2013), Zheng et al. (2014) and Guo (2011). Besides the independent and identically distributed (i.i.d) noises, chaos noise is also widely used and has been shown to be effective (Ahmed et al. 2011; Uwate et al. 2004) when injected into the gradient training process of feedforward neural networks. Chaos injection enhances the resemblance to biological systems (Li and Nara 2008; Yoshida et al. 2010), and the dynamic variation that it introduces facilitates escaping from local minima and thus improves the convergence (Ahmed et al. 2011). However, as the chaos is not an i.i.d variable, the existing convergence results and the corresponding analysis methods for noise injection-based online gradient methods can not be directly applied to the chaos injection-based gradient methods (CIBGM).

Motivated by the above issues, in this paper we try to theoretically analyze the convergence of CIBGM, covering both the batch learning and the online learning. The weak convergence and strong convergence of the algorithms will be established. The online learning we considered in this paper is the case where the training samples are fed into the network in a fixed sequence, which is also called cyclic learning in literature (Heskes and Wiegerinck 1996). Thus, compared with the convergence results for NIBOGM (Sum et al. 2012a, b; Ho et al. 2010), where the training samples are fed into the network in a totally random sequence, our results will be of deterministic nature.

The remainder of this paper is organized as follows. The network structure and CIBGM are described in "Network structure and chaos injection-based gradient method" section. "Convergence results" section presents some assumptions and our main theorems. The detailed proof of the theorems is given in "Proofs" section. In "Simulation results" section, we use a simulation example to illustrate the theoretical analysis. We conclude the paper in "Conclusion" section.

Network structure and chaos injection-based gradient method

In this section, we first introduce the network structure, which is a typical three-layer neural network. Then we describe the chaos injection-based batch gradient method and the chaos injection-based online gradient method.

Network structure

Consider a three-layer network consisting of p input nodes, q hidden nodes, and 1 output node. Let $\mathbf{w}_0 = (w_{01}, w_{02}, \ldots, w_{0q})^T \in \mathbb{R}^q$ be the weight vector between all the hidden units and the output unit, and $\mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{ip})^T \in \mathbb{R}^p$ be the weight vector between all the input units and the hidden unit $i\,(i = 1, 2, \ldots, q)$. To simplify the presentation, we write all the weight parameters in a compact form, i.e., $\mathbf{w} = \left(\mathbf{w}_0^T, \mathbf{w}_1^T, \ldots, \mathbf{w}_q^T\right)^T \in \mathbb{R}^{q+pq}$ and we define a matrix $\mathbf{V} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_q)^T \in \mathbb{R}^{q\times p}$.

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

$$\zeta = g(\mathbf{w}_0 \cdot \mathbf{F}(\mathbf{V}\boldsymbol{\xi})), \tag{1}$$

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

Chaos injection-based batch gradient method

Suppose that $\{\boldsymbol{\xi}^j, O^j\}_{j=1}^J \subset \mathbb{R}^p \times \mathbb{R}$ is a given set of training samples. The aim of the network training is to find the appropriate network weights $\mathbf{w}^*$ that can minimize the error function

$$\begin{aligned} E(\mathbf{w}) &= \frac{1}{2}\sum_{j=1}^{J}\left(O^j - g(\mathbf{w}_0 \cdot \mathbf{F}(\mathbf{V}\boldsymbol{\xi}^j))\right)^2 \\ &= \sum_{j=1}^{J} e_j\left(\mathbf{w}_0 \cdot \mathbf{F}(\mathbf{V}\boldsymbol{\xi}^j)\right), \end{aligned} \tag{2}$$

where $e_j(t) := \frac{1}{2}(O^j - g(t))^2$.

The gradient of the error function is given by

$$E_{\mathbf{w}}(\mathbf{w}) = \left(E_{\mathbf{w}_0}^T(\mathbf{w}), E_{\mathbf{w}_1}^T(\mathbf{w}), \ldots, E_{\mathbf{w}_q}^T(\mathbf{w})\right)^T \tag{3}$$

with

$$E_{\mathbf{w}_0}(\mathbf{w}) = \sum_{j=1}^{J} e_j'\left(\mathbf{w}_0 \cdot \mathbf{F}(\mathbf{V}\boldsymbol{\xi}^j)\right)\mathbf{F}(\mathbf{V}\boldsymbol{\xi}^j), \tag{4a}$$

$$E_{\mathbf{w}_i}(\mathbf{w}) = \sum_{j=1}^{J} e_j'\left(\mathbf{w}_0 \cdot \mathbf{F}(\mathbf{V}\boldsymbol{\xi}^j)\right)w_{0i}f'(\mathbf{w}_i \cdot \boldsymbol{\xi}^j)\boldsymbol{\xi}^j, \quad i = 1, 2, \ldots, q. \tag{4b}$$

Starting from an arbitrary initial value $\mathbf{w}^0$, the chaos injection-based batch gradient method updates the weights $\{\mathbf{w}^n\}$ iteratively by

$$\mathbf{w}^{n+1} = \mathbf{w}^n - \eta_n(E_{\mathbf{w}}(\mathbf{w}^n) + \eta_n A v(n)\mathbf{I}), \quad n = 0, 1, 2, \ldots, \tag{5}$$

where $\eta_n > 0$ is the learning rate, A is a positive parameter, $\mathbf{I} = (1, \ldots, 1)^T$, and

$$v(n) = \alpha v(n-1)(1 - v(n-1)) \tag{6}$$

is the logistic map/Verhust equation which is highly sensitive to the initial value $v(0)$ and the parameter α. For specific values of $v(0)$ (e.g., $0 < v(0) < 1$) and α (e.g., $3.6 < \alpha < 4$), the logistic map produces a chaotic time series.

Chaos injection-based online gradient method

The batch gradient method given in (5) updates the weights after all the training samples are fed into the network. This seems not so efficient if the training set is made up of a large number of samples. In this case, the online gradient method is preferred.

We consider the case that the training samples are supplied to the network in a fixed order in the training process. Starting from an arbitrary initial value $\mathbf{w}^0$, the chaos injection-based online gradient method updates the weights iteratively by

$$\mathbf{w}_0^{nJ+j} = \mathbf{w}_0^{nJ+j-1} + \triangle_j \mathbf{w}_0^{nJ+j-1} \tag{7a}$$

$$\mathbf{w}_i^{nJ+j} = \mathbf{w}_i^{nJ+j-1} + \triangle_j \mathbf{w}_i^{nJ+j-1}, \quad i = 1, 2, \ldots, q \tag{7b}$$

with

$$\begin{aligned} \triangle_k \mathbf{w}_0^{nJ+j-1} = &-\eta_n \Big(e_k' \Big(\mathbf{w}_0^{nJ+j-1} \cdot \mathbf{F}\big(\mathbf{V}^{nJ+j-1} \boldsymbol{\xi}^k\big) \Big) \\ &\mathbf{F}\big(\mathbf{V}^{nJ+j-1} \boldsymbol{\xi}^k\big) + \eta_n A v(n) \mathbf{I} \Big) \end{aligned} \tag{8a}$$

$$\begin{aligned} \triangle_k \mathbf{w}_i^{nJ+j-1} = &-\eta_n \Big(e_k' \Big(\mathbf{w}_0^{nJ+j-1} \cdot \mathbf{F}\big(\mathbf{V}^{nJ+j-1} \boldsymbol{\xi}^k\big) \Big) w_{0i}^{nJ+j-1} \\ &f' \Big(\mathbf{w}_i^{nJ+j-1} \cdot \boldsymbol{\xi}^k \Big) \boldsymbol{\xi}^k + \eta_n A v(n) \mathbf{I} \Big) \end{aligned} \tag{8b}$$

for $j, k = 1, 2, \ldots, J$, where $\eta_n > 0$ is the learning rate, whose value may be changed after each cycle of the training procedure, A is a positive parameter, and $v(n)$ is defined by (6).

Remark 1 During the training process, the injected chaos should be large at the beginning in order to help the gradient method avoid trapping into a local minimum, and then be smaller and smaller as the iteration (cycle) proceeds for the sake of ensuring the convergence of the algorithm to a minimum. Thus, in (5) and (8), we use $\eta_n A$ to control the magnitude of the chaos injected. Here p is used to magnify the effect of the chaos in the early training stage, and η_n [as suggested by Assumption ($A2$) in the next section, $\lim_{n\to\infty} \eta_n = 0$] is for the purpose of diminishing the effect of the injected chaos on the convergence of the algorithm with the iteration (cycle) increasing.

Convergence results

In this section, we give the convergence results of the CIBGM, covering both the batch learning case (5) and the online learning case (7).

Let $\Phi = \{\mathbf{w} : E_{\mathbf{w}}(\mathbf{w}) = 0\}$ be the stationary point set of the error function $E(\mathbf{w})$, and $\Phi_s = \{w_{ij} : \mathbf{w} = (w_{01}, \ldots, w_{ij}, \ldots, w_{qp}) \in \Phi, s = q + (i-1)p + j \quad (\text{if } i > 0) \text{ or } j \,(\text{ if } i = 0)\}$ be the projection of Φ onto the (s)th coordinate axis, for $s = 1, \ldots, pq + q$. The following assumptions are needed for our convergence results.

($A1$) $f'(t)$ and $g'(t)$ are Lipschitz continuous on any bounded closed interval;
($A2$) $\eta_n > 0, \sum_{n=0}^{\infty} \eta_n = \infty, \sum_{n=0}^{\infty} \eta_n^2 < \infty$;
($A3$) $\{\mathbf{w}^n\}$ generated by (5) is bounded over $\mathbb{R}^{pq+q}$;
($A3'$) $\{\mathbf{w}^{nJ+j}\}$ (or simply denoted by $\{\mathbf{w}^m\}$ with $m = nJ + j$) generated by (7) is bounded over $\mathbb{R}^{pq+q}$;
($A4$) The set Φ_s does not contain any interior point for every $s = 1, \ldots, pq + q$.

Remark 2 Assumption ($A1$) is satisfied by most of the activation functions, such as sigmoid functions and linear functions. Assumption ($A2$) is a traditional condition for the convergence analysis of the online gradient method (Sum et al. 2012a, b; Ho et al. 2010). Here we also use this condition in the convergence analysis of the chaos injection-based batch gradient method for the sake of controlling the impact of the chaos on the convergence of the algorithm. Assumption ($A3$) [or Assumption ($A3'$)] is a commonly used condition for the convergence analysis of the gradient method in the literature (Wu et al. 2011). In fact, this condition can be easily satisfied by adding a penalty term to the error function (Zhang et al. 2009, 2012). Assumption ($A4$) is provided to establish the strong convergence.

Now we present our convergence results, where we use "$\|\cdot\|$" to denote the Euclidean norm of a vector.

Theorem 1 *Suppose that the error function is given by* (2) *and that the weight sequence* $\{\mathbf{w}^n\}$ *is generated by the algorithm* (5) *for any initial value* $\mathbf{w}^0$. *Assume the conditions* ($A1$)–($A3$) *are valid. Then there hold the weak convergence results*

$$(a) \quad \text{There is } E^* > 0 \text{ such that } \lim_{n\to\infty} E(\mathbf{w}^n) = E^*; \tag{9}$$

$$(b) \quad \lim_{n\to\infty} \|E_{\mathbf{w}}(\mathbf{w}^n)\| = 0. \tag{10}$$

Moreover, if Assumption ($A4$) *is valid, then there holds the strong convergence, i.e., there exists a point* $\mathbf{w}^* \in \Phi$ *such that*

$$(c) \quad \lim_{n\to\infty} \mathbf{w}^n = \mathbf{w}^*. \tag{11}$$

Theorem 2 *Suppose that the conditions* ($A1$), ($A2$) *and* ($A3'$) *are valid. Then, starting from an arbitrary initial value* $\mathbf{w}^0$, *the weight sequence* $\{\mathbf{w}^m\}$ *defined by* (7) *satisfies the following weak convergence*

$$(a)\quad \textit{There is } E^{\star} > 0 \textit{ such that } \lim_{m\to\infty} E(\mathbf{w}^m) = E^{\star}; \tag{12}$$

$$(b)\quad \lim_{m\to\infty} \|E_{\mathbf{w}}(\mathbf{w}^m)\| = 0. \tag{13}$$

Moreover, if Assumption (A4) is valid, then there holds the strong convergence: there exists $\mathbf{w}^{\star} \in \Phi$ *such that*

$$(c)\quad \lim_{m\to\infty} \mathbf{w}^m = \mathbf{w}^{\star}. \tag{14}$$

Proofs

In this section, we first list several lemmas in the literature, then we conduct the proofs of theorems 1 and 2 in "Proof of Theorem 1" and "Proof of Theorem 2" subsections, respectively.

Lemma 1 (See Lemma 1 in Bertsekas and Tsitsiklis 2000) *Let* Y_n, W_n *and* Z_n *be three sequences such that* W_n *is nonnegative for all n. Assume that*

$$Y_{n+1} \le Y_n - W_n + Z_n, \quad n = 0, 1\ldots$$

and that the series $\sum_{n=0}^{\infty} Z_n$ *is convergent. Then either* $Y_n \to -\infty$ *or else* Y_n *converges to a finite value and* $\sum_{n=0}^{\infty} W_n < \infty$.

Lemma 2 (See Lemma 4.2 in Wu et al. 2011) *Suppose that the learning rate* η_n *satisfies Assumption (A2) and that the sequence* $\{a_n\}(n \in \mathbb{N})$ *satisfies* $a_n \ge 0, \sum_{n=0}^{\infty} \eta_n a_n^{\beta} < \infty$ *and* $|a_{n+1} - a_n| \le \mu\eta_n$ *for some positive constants* β *and* μ. *Then there holds* $\lim_{n\to\infty} a_n = 0$.

Lemma 3 (See Lemma 5.3 in Wang et al. 2011) *Let* $F : \Omega \subset \mathbb{R}^k \to \mathbb{R}, (k \ge 1)$ *be continuous for a bounded closed region* Ω, *and* $\Phi = \{\mathbf{z} \in \Omega : F(\mathbf{z}) = 0\}$. *The projection of* Φ *on each coordinate axis does not contain any interior point. Let the sequence* $\{\mathbf{z}^n\}$ *satisfy:*

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

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

Proof of Theorem 1

Lemma 4 *Suppose the conditions (A1) and (A3) are valid, then* $E_{\mathbf{w}}(\mathbf{w})$ *satisfies Lipschitz conditon, that is, there exists a positive constant L, such that*

$$\left\|E_{\mathbf{w}}(\mathbf{w}^{n+1}) - E_{\mathbf{w}}(\mathbf{w}^n)\right\| \le L\left\|\mathbf{w}^{n+1} - \mathbf{w}^n\right\|. \tag{15}$$

Specially, for $\theta \in [0, 1]$, *there holds*

$$\left\|E_{\mathbf{w}}(\mathbf{w}^n + \theta(\mathbf{w}^{n+1} - \mathbf{w}^n)) - E_{\mathbf{w}}(\mathbf{w}^n)\right\| \le L\theta\|\mathbf{w}^{n+1} - \mathbf{w}^n\|. \tag{16}$$

Proof The proof of this lemma is similar to Lemma 2 of Zhang et al. (2012) and thus omitted. □

Proof of (9) Given that $0 < v(0) < 1$ and $3.6 < \alpha < 4$, it is easy to see

$$\begin{aligned} 0 < v(n) &= \alpha v(n-1)(1 - v(n-1)) \\ &\le \alpha \frac{(v(n-1) + 1 - v(n-1))^2}{4} = \frac{\alpha}{4} < 1. \end{aligned} \tag{17}$$

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

$$\begin{aligned} &E(\mathbf{w}^{n+1}) - E(\mathbf{w}^n) \\ &= \left(E_{\mathbf{w}}(\mathbf{w}^n + \theta(\mathbf{w}^{n+1} - \mathbf{w}^n))\right)^T(\mathbf{w}^{n+1} - \mathbf{w}^n) \\ &= (E_{\mathbf{w}}(\mathbf{w}^n))^T(\mathbf{w}^{n+1} - \mathbf{w}^n) \\ &\quad + \left(E_{\mathbf{w}}(\mathbf{w}^n + \theta(\mathbf{w}^{n+1} - \mathbf{w}^n)) - (E_{\mathbf{w}}(\mathbf{w}^n))\right)^T(\mathbf{w}^{n+1} - \mathbf{w}^n) \\ &\le (E_{\mathbf{w}}(\mathbf{w}^n))^T(\mathbf{w}^{n+1} - \mathbf{w}^n) + L\theta\|\mathbf{w}^{n+1} - \mathbf{w}^n\|^2, \end{aligned} \tag{18}$$

where the last inequality is due to (16). Considering (5) and (18), we have

$$\begin{aligned} E(\mathbf{w}^{n+1}) \le E(\mathbf{w}^n) &+ \eta_n(E_{\mathbf{w}}(\mathbf{w}^n))^T(-E_{\mathbf{w}}(\mathbf{w}^n) - \eta_n A v(n)\mathbf{I}) \\ &+ L\theta\eta_n\|E_{\mathbf{w}}(\mathbf{w}^n) + A\eta_n v(n)\mathbf{I}\|^2. \end{aligned} \tag{19}$$

Using (17) and the inequality $\|E_{\mathbf{w}}(\mathbf{w}^n)\| \le \frac{(1+\|E_{\mathbf{w}}(\mathbf{w}^n)\|^2)}{2}$, the second term on the right hand side of (19) can be evaluated

$$\begin{aligned} &(E_{\mathbf{w}}(\mathbf{w}^n))^T[-E_{\mathbf{w}}(\mathbf{w}^n) - \eta_n A v(n)\mathbf{I}] \\ &\le -\|(E_{\mathbf{w}}(\mathbf{w}^n))\|^2 + \eta_n A\sqrt{pq+q}\|E_{\mathbf{w}}(\mathbf{w}^n)\| \\ &= -\|(E_{\mathbf{w}}(\mathbf{w}^n))\|^2 + \eta_n \frac{A}{2}\sqrt{pq+q}\left(1 + \|E_{\mathbf{w}}(\mathbf{w}^n)\|^2\right). \end{aligned} \tag{20}$$

Using inequality $(a+b)^2 \le 2(a^2+b^2)$, the third term on the right hand side of (19) can be evaluated

$$\begin{aligned} &\left\|\eta_n E_{\mathbf{w}}(\mathbf{w}^n) + A\eta_n^2 v(n)\mathbf{I}\right\|^2 \\ &\le 2\eta_n^2\|E_{\mathbf{w}}(\mathbf{w}^n)\|^2 + 2\eta_n^4 A^2\|\mathbf{I}\|^2 \\ &\le 2\eta_n^2\|E_{\mathbf{w}}(\mathbf{w}^n)\|^2 + 2A^2(pq+q)\eta_n^4. \end{aligned} \tag{21}$$

Combining (19)–(21), we have

$$\begin{aligned} E(\mathbf{w}^{n+1}) &\le E(\mathbf{w}^n) - \eta_n\|E_{\mathbf{w}}(\mathbf{w}^n)\|^2 \\ &\quad + \eta_n^2 \frac{A}{2}\sqrt{pq+q}\left(1 + \|E_{\mathbf{w}}(\mathbf{w}^n)\|^2\right) \\ &\quad + 2L\theta\eta_n^2\|E_{\mathbf{w}}(\mathbf{w}^n)\|^2 + 2L\theta A^2(pq+q)\eta_n^4 \\ &= E(\mathbf{w}^n) - \eta_n\|E_{\mathbf{w}}(\mathbf{w}^n)\|^2 \\ &\quad + \eta_n^2\left(\frac{A}{2}\sqrt{pq+q} + 2L\theta A^2\eta_n^2(pq+q)\right. \\ &\quad \left. + \left(2L\theta + \frac{A}{2}\sqrt{pq+q}\right)\|E_{\mathbf{w}}(\mathbf{w}^n)\|^2\right). \end{aligned} \tag{22}$$

By Assumptions (A1) and (A3), there is a constant $C_1 > 0$ such that for all $n = 0, 1, \ldots$

$$\|E_{\mathbf{w}}(\mathbf{w}^n)\| \le C_1. \tag{23}$$

Thus, there exists a positive constant C_2, such that

$$E(\mathbf{w}^{n+1}) \le E(\mathbf{w}^n) - \eta_n \|E_{\mathbf{w}}(\mathbf{w}^n))\|^2 + \eta_n^2 C_2. \tag{24}$$

Combining $\sum_{n=1}^{\infty} \eta_n^2 C_2 < \infty, E(\mathbf{w}^n) > 0$, and according to Lemma 1, we can conclude that there exists a constant E^* such that

$$\lim_{n\to\infty} E(\mathbf{w}^n) = E^* \tag{25}$$

and

$$\sum_{n=0}^{\infty} \|E_{\mathbf{w}}(\mathbf{w}^n)\|^2 \eta_n < \infty. \tag{26}$$

This completes the proof of (9). □

Proof of (10) Using (5), (15) and (23), we have

$$\begin{aligned} \left| \|E_{\mathbf{w}}(\mathbf{w}^{n+1})\| - \|E_{\mathbf{w}}(\mathbf{w}^n)\| \right| &\le \|E_{\mathbf{w}}(\mathbf{w}^{n+1}) - E_{\mathbf{w}}(\mathbf{w}^n)\| \\ &\le L\|\mathbf{w}^{n+1} - \mathbf{w}^n\| \\ &\le \eta_n L(\|E_{\mathbf{w}}(\mathbf{w}^n)\| + \eta_n A\|\mathbf{I}\|) \\ &\le C_3 \eta_n, \end{aligned} \tag{27}$$

where $C_3 = L(C_1 + A\sqrt{pq+q}\sup_{n\in\mathbb{N}} \eta_n)$. Thus, by (26), (27), and Lemma 2, we conclude

$$\lim_{n\to\infty} E_{\mathbf{w}}(\mathbf{w}^n) = 0. \qquad \square$$

Proof of (11) Obviously $\|E_{\mathbf{w}}(\mathbf{w})\|$ is a continuous function under the Assumption (A1). Using (5), (17) and (23), we have

$$\lim_{n\to\infty} \|\mathbf{w}^{n+1} - \mathbf{w}^n\| = \lim_{n\to\infty} \eta_n \|E_{\mathbf{w}}(\mathbf{w}^n) + A\eta_n v(n)\mathbf{I}\| = 0. \tag{28}$$

Furthermore, the Assumption (A4) is valid. Thus, applying Lemma 3, there exists a unique $\mathbf{w}^* \in \Phi$ such that $\lim_{n\to\infty} \mathbf{w}^n = \mathbf{w}^*$. □

Proof of Theorem 2

Let the sequence $\{\mathbf{w}^{nJ+j}\}(n \in \mathbb{N}, j = 1, 2, \ldots, J)$ be generated by (7). For brevity, we introduce the following notations:

$$\mathbf{F}^{nJ+j,k} = \mathbf{F}(\mathbf{V}^{nJ+j}\boldsymbol{\xi}^k), \tag{29a}$$

$$\mathbf{r}_i^{n,j} = \triangle_j \mathbf{w}_i^{nJ+j-1} - \triangle_j \mathbf{w}_i^{nJ}, \tag{29b}$$

$$\mathbf{h}_i^{n,l} = \mathbf{w}_i^{nJ+l} - \mathbf{w}_i^{nJ} = \sum_{j=1}^{l} \triangle_j \mathbf{w}_i^{nJ+j-1} = \sum_{j=1}^{l} \triangle_j \mathbf{w}_i^{nJ} + \sum_{j=1}^{l} \mathbf{r}_i^{n,j}, \tag{29c}$$

$$\boldsymbol{\psi}^{n,l,j} = \mathbf{F}^{nJ+l,j} - \mathbf{F}^{nJ,j}, \tag{29d}$$

for $n \in \mathbb{N}; j, k, l = 1, 2, \ldots, J; i = 0, 1, 2, \ldots, q$.

Lemma 5 (See Lemma 4.1 in Wu et al. 2011) *Let $h(x)$ be a function defined on a bounded closed interval $[a, b]$ such that $h'(x)$ is Lipschitz continuous with Lipschitz constant $K > 0$. Then, $h'(x)$ is differentiable almost everywhere in $[a, b]$ and*

$$|h''(x)| \le K, \quad x \in [a, b]. \tag{30}$$

Moreover, there exists a constant $C_4 > 0$ such that

$$\begin{aligned} h(x) \le h(x_0) + h'(x_0)(x - x_0) + C_4(x - x_0)^2, \\ \forall x_0, \quad x \in [a, b]. \end{aligned} \tag{31}$$

Lemma 6 *Suppose the conditions (A1) and (A3′) are valid, and the sequence $\{\mathbf{w}^{nJ+j}\}$ is generated by (7). Then there are $C_5 - C_8$ such that*

$$\|\mathbf{F}^{nJ+j,k}\| \le C_5, \tag{32}$$

$$\|\mathbf{h}_i^{n,l}\| \le C_6 \eta_n, \tag{33}$$

$$\|\boldsymbol{\psi}^{n,l,j}\| \le C_7 \eta_n, \tag{34}$$

$$\|\mathbf{r}_i^{n,j}\| \le C_8 \eta_n^2, \tag{35}$$

where $n \in \mathbb{N}; j, k, l = 1, 2, \ldots, J; i = 0, 1, 2, \ldots, q$.

Proof According to Assumption (A3′), we can define a constant $C_w = \sup \|\mathbf{w}^m\|$. Then we have

$$\left|\mathbf{w}_i^{nJ+j} \cdot \boldsymbol{\xi}^k\right| \le \left\|\mathbf{w}_i^{nJ+j}\right\| \left\|\boldsymbol{\xi}^k\right\| \le C_w \max_{1\le k\le J} \|\boldsymbol{\xi}^k\| = C_9. \tag{36}$$

Accordingly, there exist two positive constants C_f and $C_{f'}$ such that

$$\sup_{|t|\le C_9} |f(t)| = C_f, \quad \sup_{|t|\le 2C_9} |f'(t)| = C_{f'}. \tag{37}$$

Thus we have

$$\|\mathbf{F}^{nJ+j,k}\| = \|\mathbf{F}(\mathbf{V}^{nJ+j}\boldsymbol{\xi}^k)\| \le \sqrt{q} C_f = C_5, \tag{38}$$

and

$$\left|\mathbf{w}_0^{nJ+j} \cdot \mathbf{F}^{nJ+j,k}\right| \le \left\|\mathbf{w}_0^{nJ+j}\right\| \left\|\mathbf{F}^{nJ+j,k}\right\| \le C_w C_5. \tag{39}$$

Then, there is a positive constant $C_{e_j'}$ such that

$$\max_{|t|\le C_w C_5} \left|e_j'(t)\right| \le C_{e_j'}. \tag{40}$$

Using (8), (17), (29c), (38) and (40), we have

$$\begin{aligned}\left\|\mathbf{h}_0^{n,l}\right\| &= \left\|\sum_{j=1}^{l}\triangle_j \mathbf{w}_0^{nJ+j-1}\right\| \\ &= \left\|-\eta_n\sum_{j=1}^{l}\left(e_j'(\mathbf{w}_0^{nJ+j-1}\cdot\mathbf{F}^{nJ+j-1,j})\right)\mathbf{F}^{nJ+j-1,j}+\eta_n Av(n)\mathbf{I}\right)\right\| \\ &\le \eta_n J C_{e_j'} C_5 + \eta_n^2 AJ\sqrt{q} \\ &\le J\left(C_{e_j'}C_5 + A\sqrt{q}\sup\eta_n\right)\eta_n.\end{aligned} \tag{41}$$

Similarly, for $i = 1,\ldots,q$, we have

$$\begin{aligned}\left\|\mathbf{h}_i^{n,l}\right\| &= \left\|\sum_{j=1}^{l}\triangle_j \mathbf{w}_i^{nJ+j-1}\right\| \\ &= \left\|-\eta_n\sum_{j=1}^{l}\left(e_j'\left(\mathbf{w}_0^{nJ+j-1}\cdot\mathbf{F}^{nJ+j-1,j}\right)w_{0i}^{nJ+j-1}\right.\right. \\ &\quad \left.\left. f'\left(\mathbf{w}_i^{nJ+j-1}\cdot\boldsymbol{\xi}^j\right)\boldsymbol{\xi}^j + \eta_n Av(n)\mathbf{I}\right)\right\| \\ &\le \eta_n J C_{e_j'} C_{f'} C_9 + \eta_n^2 AJ\sqrt{p} \\ &\le J\left(C_{e_j'}C_{f'}C_9 + A\sqrt{p}\sup\eta_n\right)\eta_n.\end{aligned}$$

Let
$C_6 = J\max\{C_{e_j'}C_5 + A\sqrt{q}\sup\eta_n, C_{e_j'}C_{f'}C_9 + A\sqrt{p}\sup\eta_n\}$, then we have $\|\mathbf{h}_i^{n,l}\| \le C_6\eta_n$ for $i = 0,1,\ldots,q$.

Using (29c), (29d), (33) and the mean value theorem, we have

$$\begin{aligned}\|\boldsymbol{\psi}^{n,l,j}\| &= \|\mathbf{F}(\mathbf{V}^{nJ+l}\boldsymbol{\xi}^j) - \mathbf{F}(\mathbf{V}^{nJ}\boldsymbol{\xi}^j)\| \\ &= \left[\sum_{i=1}^{q}\left[f\left(\mathbf{w}_i^{nJ+l}\cdot\boldsymbol{\xi}^j\right) - f\left(\mathbf{w}_i^{nJ}\cdot\boldsymbol{\xi}^j\right)\right]^2\right]^{\frac{1}{2}} \\ &= \left[\sum_{i=1}^{q}\left[f'\left(\mathbf{w}_i^{nJ}\cdot\boldsymbol{\xi}^j + \theta_i\left(\mathbf{w}_i^{nJ+l}\cdot\boldsymbol{\xi}^j - \mathbf{w}_i^{nJ}\cdot\boldsymbol{\xi}^j\right)\right)\mathbf{h}_i^{n,l}\cdot\boldsymbol{\xi}^j\right]^2\right]^{\frac{1}{2}} \\ &\le C_{f'}\|\boldsymbol{\xi}^j\|\sum_{i=1}^{q}\|\mathbf{h}_i^{n,l}\| \le qC_{f'}\max_{1\le j\le J}\|\boldsymbol{\xi}^j\|C_6\eta_n = C_7\eta_n,\end{aligned} \tag{42}$$

where $\theta_i \in (0,1)$ and $C_7 = qC_{f'}\max_{1\le j\le J}\|\boldsymbol{\xi}^j\|C_6$.

As Assumptions $(A1)$ and $(A3')$ are valid, it is easy to see that there exists a constant L' such that for any $n = 0,1,\ldots,$ and $1 \le k_1,k_2,j_1,j_2,l_1,l_2 \le J$, there holds

$$\begin{aligned}&\left|e_j'\left(\mathbf{w}_0^{nJ+k_1}\cdot\mathbf{F}^{nJ+j_1,l_1}\right) - e_j'\left(\mathbf{w}_0^{nJ+k_2}\cdot\mathbf{F}^{nJ+j_2,l_2}\right)\right| \\ &\le L'\left|\mathbf{w}_0^{nJ+k_1}\cdot\mathbf{F}^{nJ+j_1,l_1} - \mathbf{w}_0^{nJ+k_2}\cdot\mathbf{F}^{nJ+j_2,l_2}\right|.\end{aligned} \tag{43}$$

Combining (8), (29), (32)–(34) and (43), we have

$$\begin{aligned}\|\mathbf{r}_0^{n,j}\| &= \left\|\triangle_j\mathbf{w}_0^{nJ+j-1} - \triangle_j\mathbf{w}_0^{nJ}\right\| \\ &= \left\|-\eta_n\left(e_j'\left(\mathbf{w}_0^{nJ+j-1}\cdot\mathbf{F}^{nJ+j-1,j}\right)\mathbf{F}^{nJ+j-1,j}\right.\right. \\ &\quad \left.\left. - e_j'(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j})\mathbf{F}^{nJ,j}\right)\right\| \\ &= \left\|-\eta_n\left[e_j'\left(\mathbf{w}_0^{nJ+j-1}\cdot\mathbf{F}^{nJ+j-1,j}\right)\boldsymbol{\psi}^{n,j-1,j}\right.\right. \\ &\quad + \left(e_j'\left(\mathbf{w}_0^{nJ+j-1}\cdot\mathbf{F}^{nJ+j-1,j}\right) - e_j'\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ+j-1,j}\right)\right)\mathbf{F}^{nJ,j} \\ &\quad \left.\left. + \left(e_j'(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ+j-1,j}) - e_j'(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j})\right)\mathbf{F}^{nJ,j}\right]\right\| \\ &\le \eta_n\left(\left|e_j'\left(\mathbf{w}_0^{nJ+j-1}\cdot\mathbf{F}^{nJ+j-1,j}\right)\right|\|\boldsymbol{\psi}^{n,j-1,j}\|\right. \\ &\quad + \left|e_j'\left(\mathbf{w}_0^{nJ+j-1}\cdot\mathbf{F}^{nJ+j-1,j}\right) - e_j'\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ+j-1,j}\right)\right|\|\mathbf{F}^{nJ,j}\| \\ &\quad \left. + \left|e_j'(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ+j-1,j}) - e_j'(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j})\right|\|\mathbf{F}^{nJ,j}\|\right) \\ &\le \eta_n\left(C_{e_j'}\|\boldsymbol{\psi}^{n,j-1,j}\| + L'\left|\mathbf{w}_0^{nJ+j-1}\right.\right. \\ &\quad \times \mathbf{F}^{nJ+j-1,j} - \mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ+j-1,j}\|\mathbf{F}^{nJ,j}\| \\ &\quad \left.\left. + L'\left|\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ+j-1,j} - \mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j}\right|\|\mathbf{F}^{nJ,j}\|\right) \\ &\le \eta_n\left(C_{e_j'}\|\boldsymbol{\psi}^{n,j-1,j}\| + L'C_5^2\left\|\mathbf{h}_0^{n,j-1}\right\| + L'C_wC_5\|\boldsymbol{\psi}^{n,j-1,j}\|\right) \\ &\le \eta_n^2\left(C_{e_j'}C_7 + L'C_5^2C_6 + L'C_wC_5C_7\right) = C_{10}\eta_n^2,\end{aligned} \tag{44}$$

where $C_{10} = C_{e_j'}C_7 + L'C_5^2C_6 + L'C_wC_5C_7$. Similarly, we can show the existence of a constant $C_{11} > 0$ such that

$$\|\mathbf{r}_i^{n,j}\| \le C_{11}\eta_n^2. \tag{45}$$

Let $C_8 = \max\{C_{10}, C_{11}\}$, then we have $\|\mathbf{r}_i^{n,j}\| \le C_8\eta_n^2$ for $i = 0,1,2,\ldots,q$. □

Lemma 7 *Let the sequence* $\{\mathbf{w}^{nJ+j}\}$ *be generated by* (7). *Under assumptions* $(A1)$ *and* $(A3')$, *there holds*

$$E(\mathbf{w}^{(n+1)J}) \le E(\mathbf{w}^{nJ}) - \eta_n\|E_{\mathbf{w}}(\mathbf{w}^{nJ})\|^2 + C_{12}\eta_n^2, \quad (n = 0,1,\ldots) \tag{46}$$

where $C_{12} > 0$ *is a positive constant.*

Proof By virtue of Assumption $(A1)$ and Lemma 5, we know that $f''(\mathbf{w}_i^{nJ}\cdot\boldsymbol{\xi}^j + t(\mathbf{h}_i^{n,J}\cdot\boldsymbol{\xi}^j))$ is integrable almost everywhere on $t \in [0,1]$. Then, using Taylor's mean value theorem we arrive at

$$\begin{aligned}\mathbf{w}_0^{nJ}\cdot\boldsymbol{\psi}^{n,J,j} &= \sum_{i=1}^{q}w_{0i}^{nJ}\left[f\left(\mathbf{w}_i^{(n+1)J}\cdot\boldsymbol{\xi}^j\right) - f\left(\mathbf{w}_i^{nJ}\cdot\boldsymbol{\xi}^j\right)\right] \\ &= \sum_{i=1}^{q}w_{0i}^{nJ}f'\left(\mathbf{w}_i^{nJ}\cdot\boldsymbol{\xi}^j\right)\mathbf{h}_i^{n,J}\cdot\boldsymbol{\xi}^j \\ &\quad + \sum_{i=1}^{q}w_{0i}^{nJ}\left(\mathbf{h}_i^{n,J}\cdot\boldsymbol{\xi}^j\right)^2 \\ &\quad \times\int_0^1(1-t)f''\left(\mathbf{w}_i^{nJ}\cdot\boldsymbol{\xi}^j + t\left(\mathbf{h}_i^{n,J}\cdot\boldsymbol{\xi}^j\right)\right)dt.\end{aligned} \tag{47}$$

By virtue of (8), (29), (47) and Lemma 5, there is a constant $C_{13} > 0$ such that

$$\begin{aligned}
&e_j\left(\mathbf{w}_0^{(n+1)J}\cdot\mathbf{F}^{(n+1)J,j}\right)\\
&\le e_j\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j}\right)+e_j'\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j}\right)\\
&\quad\left(\mathbf{w}_0^{(n+1)J}\cdot\mathbf{F}^{(n+1)J,j}-\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j}\right)\\
&\quad+C_{13}\left(\mathbf{w}_0^{(n+1)J}\cdot\mathbf{F}^{(n+1)J,j}-\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j}\right)^2\\
&=e_j\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j}\right)+e_j'\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j}\right)\\
&\quad\left(\mathbf{h}_0^{n,J}\cdot\mathbf{F}^{nJ,j}+\mathbf{w}_0^{nJ}\cdot\boldsymbol{\psi}^{n,J,j}+\mathbf{h}_0^{n,J}\cdot\boldsymbol{\psi}^{n,J,j}\right)\\
&\quad+C_{13}\left(\mathbf{h}_0^{n,J}\cdot\mathbf{F}^{nJ,j}+\mathbf{w}_0^{nJ}\cdot\boldsymbol{\psi}^{n,J,j}+\mathbf{h}_0^{n,J}\cdot\boldsymbol{\psi}^{n,J,j}\right)^2\\
&=e_j\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j}\right)+e_j'\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j}\right)\mathbf{F}^{nJ,j}\cdot\mathbf{h}_0^{n,J}\\
&\quad+e_j'\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j}\right)\sum_{i=1}^{q}w_{0i}^{nJ}f'\left(\mathbf{w}_i^{nJ}\cdot\xi^j\right)\xi^j\cdot\mathbf{h}_i^{n,J}+\delta_1\\
&=e_j\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j}\right)+e_j'\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j}\right)\mathbf{F}^{nJ,j}\\
&\quad\cdot\left(-\eta_n\sum_{k=1}^{J}\left[e_k'\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,k}\right)\mathbf{F}^{nJ,k}+\eta_nAv(n)\mathbf{I}\right]+\sum_{k=1}^{J}\mathbf{r}_0^{n,k}\right)\\
&\quad+e_j'\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j}\right)\sum_{i=1}^{q}w_{0i}^{nJ}f'\left(\mathbf{w}_i^{nJ}\cdot\xi^j\right)\xi^j\cdot\\
&\quad\left(-\eta_n\sum_{k=1}^{J}\left[e_k'\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,k}\right)w_{0i}^{nJ}f'\left(\mathbf{w}_i^{nJ}\cdot\xi^k\right)\xi^k\right.\right.\\
&\left.\left.+\eta_nAv(n)\mathbf{I}\right]+\sum_{k=1}^{J}\mathbf{r}_i^{n,k}\right)+\delta_1=e_j\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j}\right)\\
&-\eta_ne_j'\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j}\right)\mathbf{F}^{nJ,j}\cdot\sum_{k=1}^{J}e_k'\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,k}\right)\mathbf{F}^{nJ,k}\\
&-\eta_ne_j'\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j}\right)\sum_{i=1}^{q}w_{0i}^{nJ}f'\left(\mathbf{w}_i^{nJ}\cdot\xi^j\right)\xi^j\\
&\times\sum_{k=1}^{J}e_k'\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,k}\right)w_{0i}^{nJ}f'\left(\mathbf{w}_i^{nJ}\cdot\xi^k\right)\xi^k+\delta_2,
\end{aligned}\tag{48}$$

where

$$\begin{aligned}
\delta_1&=e_j'\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j}\right)\sum_{i=1}^{q}w_{0i}^{nJ}\left(\mathbf{h}_i^{n,J}\cdot\xi^j\right)^2\\
&\quad\times\int_0^1(1-t)f''\left(\mathbf{w}_i^{nJ}\cdot\xi^j+t\left(\mathbf{h}_i^{n,J}\cdot\xi^j\right)\right)dt\\
&\quad+e_j'\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j}\right)\mathbf{h}_0^{n,J}\cdot\boldsymbol{\psi}^{n,J,j}\\
&\quad+C_{13}\left(\mathbf{h}_0^{n,J}\cdot\mathbf{F}^{nJ,j}+\mathbf{w}_0^{nJ}\cdot\boldsymbol{\psi}^{n,J,j}+\mathbf{h}_0^{n,J}\cdot\boldsymbol{\psi}^{n,J,j}\right)^2
\end{aligned}\tag{49}$$

and

$$\begin{aligned}
\delta_2&=e_j'\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j}\right)\mathbf{F}^{nJ,j}\cdot\left(-\eta_n^2JAv(n)\mathbf{I}+\sum_{k=1}^{J}\mathbf{r}_0^{n,k}\right)\\
&\quad+e_j'\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j}\right)\sum_{i=1}^{q}w_{0i}^{nJ}f'\left(\mathbf{w}_i^{nJ}\cdot\xi^j\right)\xi^j\\
&\quad\times\left(-\eta_n^2JAv(n)\mathbf{I}+\sum_{k=1}^{J}\mathbf{r}_i^{n,k}\right)+\delta_1.
\end{aligned}\tag{50}$$

Summing (48) for j from 1 to J up, and noticing (2), (4a, 4b), (29a), (49) and (50), we have

$$\begin{aligned}
E(\mathbf{w}^{(n+1)J})&\le E(\mathbf{w}^{nJ})-\eta_n\sum_{i=0}^{q}\left\|E_{\mathbf{w}_i}(\mathbf{w}^{nJ})\right\|^2+\delta_3\\
&=E(\mathbf{w}^{nJ})-\eta_n\left\|E_{\mathbf{w}}(\mathbf{w}^{nJ})\right\|^2+\delta_3,
\end{aligned}\tag{51}$$

where

$$\begin{aligned}
\delta_3&=E_{\mathbf{w}_0}(\mathbf{w}^{nJ})\cdot\left(-\eta_n^2JAv(n)\mathbf{I}+\sum_{k=1}^{J}\mathbf{r}_0^{n,k}\right)\\
&\quad+\sum_{i=1}^{q}E_{\mathbf{w}_i}(\mathbf{w}^{nJ})\cdot\left(-\eta_n^2JAv(n)\mathbf{I}+\sum_{k=1}^{J}\mathbf{r}_i^{n,k}\right)\\
&\quad+\sum_{j=1}^{J}e_j'\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j}\right)\sum_{i=1}^{q}w_{0i}^{nJ}\left(\mathbf{h}_i^{n,J}\cdot\xi^j\right)^2\\
&\quad\times\int_0^1(1-t)f''\left(\mathbf{w}_i^{nJ}\cdot\xi^j+t\left(\mathbf{h}_i^{n,J}\cdot\xi^j\right)\right)dt\\
&\quad+\sum_{j=1}^{J}e_j'\left(\mathbf{w}_0^{nJ}\cdot\mathbf{F}^{nJ,j}\right)\mathbf{h}_0^{n,J}\cdot\boldsymbol{\psi}^{n,J,j}\\
&\quad+C_{13}\sum_{j=1}^{J}\left(\mathbf{h}_0^{n,J}\cdot\mathbf{F}^{nJ,j}+\mathbf{w}_0^{nJ}\cdot\boldsymbol{\psi}^{n,J,j}+\mathbf{h}_0^{n,J}\cdot\boldsymbol{\psi}^{n,J,j}\right)^2.
\end{aligned}\tag{52}$$

Considering (17) and (32)–(40), it is easy to see that there exists a constant C_{12} such that

$$\delta_3<C_{12}\eta_n^2.\tag{53}$$

Thus, the desired estimate is deduced by combining (51) and (53). □

Now we are ready to prove the convergence theorem.

Proof of (12) According to Lemmas 1 and 7, there exists a constant $E^\star$ such that $\lim_{m\to\infty}E(\mathbf{w}^m)=E^\star$ or $\lim_{m\to\infty}E(\mathbf{w}^m)=-\infty$. Recall $E(\mathbf{w}^m)\geqslant 0$, then we have (12). □

Proof of (13) Using Lemmas 1 and 7, we have that

$$\sum_{n=0}^{\infty}\eta_n\left\|E_{\mathbf{w}}(\mathbf{w}^{nJ})\right\|^2<\infty.\tag{54}$$

Similarly to Lemma 4, there exists a Lipschitz constant L'' such that

$$\left\|E_{\mathbf{w}}(\mathbf{w}^{m+l})-E_{\mathbf{w}}(\mathbf{w}^m)\right\|\le L''\left\|\mathbf{w}^{m+l}-\mathbf{w}^m\right\|.\tag{55}$$

where $\mathbf{w}^m$ is the weight sequence generated by (7) and l is a positive integer.

Using (55) and (33), for $n=0,1,\ldots$, and $j=1,\ldots,J$, we have

$$\begin{aligned}\left|\left\|E_{\mathbf{w}}(\mathbf{w}^{(n+1)J})\right\| - \left\|E_{\mathbf{w}}(\mathbf{w}^{nJ})\right\|\right| &\le \left\|E_{\mathbf{w}}(\mathbf{w}^{(n+1)J}) - E_{\mathbf{w}}(\mathbf{w}^{nJ})\right\| \\ &\le L''\|\mathbf{w}^{(n+1)J} - \mathbf{w}^{nJ}\| \\ &= L''\sqrt{\sum_{i=0}^{q}\left\|\mathbf{h}_i^{nJ}\right\|^2} \\ &\le \sqrt{q+1}L''C_6\eta_n. \end{aligned} \tag{56}$$

Combining (54), (56) and Lemma 2, we have

$$\lim_{n\to\infty}\left\|E_{\mathbf{w}}(\mathbf{w}^{nJ})\right\| = 0. \tag{57}$$

Since

$$\begin{aligned}\left\|E_{\mathbf{w}}(\mathbf{w}^{nJ+j})\right\| &\le \left\|E_{\mathbf{w}}(\mathbf{w}^{nJ+j}) - E_{\mathbf{w}}(\mathbf{w}^{nJ})\right\| + \left\|E_{\mathbf{w}}(\mathbf{w}^{nJ})\right\| \\ &\le L''C_6\sqrt{q+1}\eta_n + \left\|E_{\mathbf{w}}(\mathbf{w}^{nJ})\right\|, \end{aligned} \tag{58}$$

we have $\lim_{n\to\infty}\|E_{\mathbf{w}}(\mathbf{w}^{nJ+j})\| = 0$ for $j = 1, 2, \ldots, J$. □

Proof of (14) The proof is almost the same as the proof of (11) and thus it is omitted here. □

Simulation results

In this section, we illustrate the convergence behavior of the CIBGM using the sonar signal classification problem.

Sonar signal classification is one of the benchmark problems in neural network field. Our task is to train a network to discriminate between sonar returns bounced off a metal cylinder and those bounced off a roughly cylindrical rock. We obtained the data set from UCI machine learning repository (http://archive.ics.uci.edu/ml/), which comprises 208 samples, each with 60 components. In this simulation, we stochastically choose 164 samples for training and 44 samples for test.

The network for training is with the structure of 60–25–2. The activation functions for both the hidden and output layers are set to be $logsig(\cdot)$ in MATLAB, which is a commonly used sigmoid function. We choose the initial weights to be random numbers in the interval $[-0.5, 0.5]$.

The simulation is carried out by choosing the parameters $A = 500, v(0) = 0.5$ and $\alpha = 3.8$. We set the learning rate $\eta_n = \frac{0.08}{164}$ if $n \le 120$ and $\frac{n^{-0.5}}{164}$ if $n > 120$, which satisfies Assumption $(A2)$. Here 164 is the number of the training samples. The maximum training iteration (cycle) is 2,000. The learning curves for the chaos injection-based batch gradient method are depicted in Fig. 1, which shows the training error tends to a constant and the gradient of the error function tends to zero. This supports our theoretical analysis. The learning curves for the chaos injection-based online gradient method are almost the same with Fig. 1, with the only change that we should replace the x label "Number of iterations" with "Number of cycles".

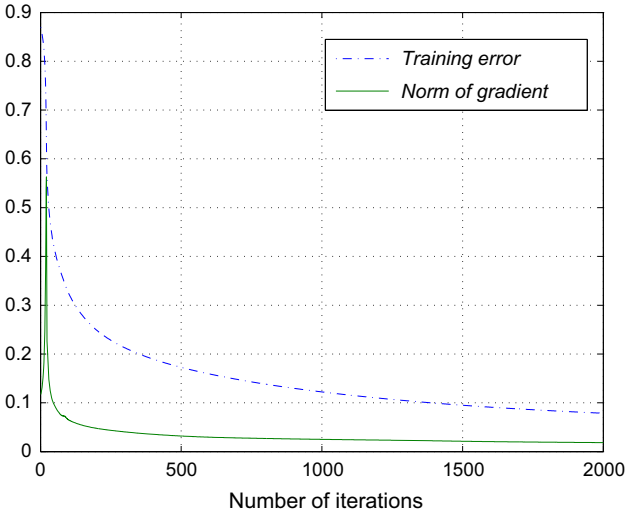

Fig. 1 Learning curves for chaos injection-based batch gradient method

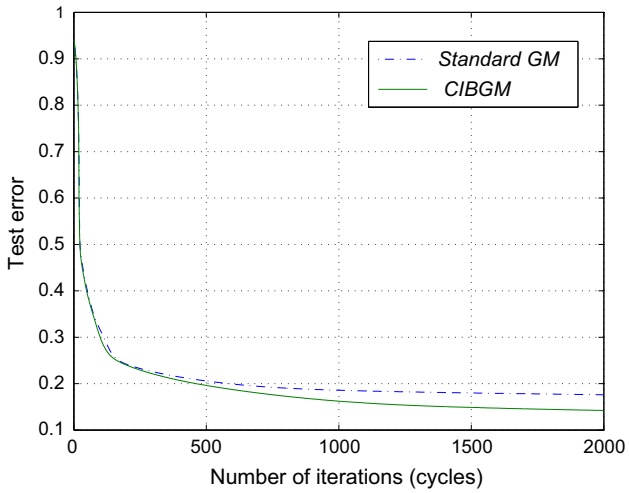

Fig. 2 Test error curves for CIBGM and GM (with no chaos injected)

In order to show the effectiveness of the chaos injection method, we compare the test error curves of CIBGM and the standard GM (with no chaos injected) in Fig. 2. We can see that the test error of CIBGM converges faster and tends to a smaller number than that of the standard GM.

We mention that, though there is no restriction for the parameter A in Theorems 1 and 2, the choice of A is still of great importance. If A is too small, CIBGM will reduce to the standard GM. On the other hand, if A is too large, then the chaos term will dominate the update of the CIBGM

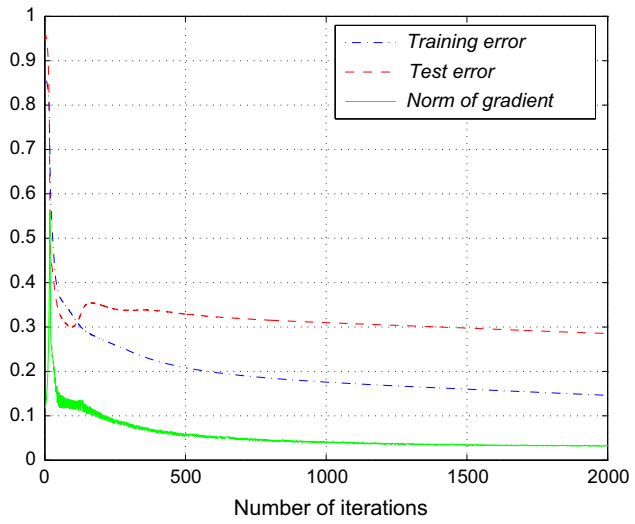

Fig. 3 Performance for chaos injection-based batch gradient method with $A = 5{,}000$

especially in the early stage of the training procedure. As a result, the algorithm will converge very slowly and the performance may even be unacceptable. Figure 3 shows the results of the chaos injection-based batch gradient method for $A = 5{,}000$. We can find that the algorithm still converges. However, the performance is much worse than that for $A = 500$.

Conclusion

This paper investigates the chaos injection-based gradient method (CIBGM) for the feedforward neural networks. Two learning mode cases, batch learning and online learning, are considered. Under the conditions that the derivatives of activation functions are Lipschitz continuous on any bounded closed interval and the learning rate η_n is positive and satisfies $\sum_{n=0}^{\infty} \eta_n = \infty$ and $\sum_{n=0}^{\infty} \eta_n^2 < \infty$, we derive the weak convergence of the CIBGM, that is the gradient of the error function tends to zero and the error function tends to a constant. The strong convergence is also derived with the assumption that the set Φ_s does not contain any interior point. The theoretical findings and the effectiveness of the CIBGM are illustrated by a simulation example. Future research includes the study on the convergence of the chaos injection-based stochastic gradient method.

Acknowledgments This work is partly supported by the National Natural Science Foundation of China (Nos. 61101228, 61301202, 61402071), the China Postdoctoral Science Foundation (No. 2012M520623), and the Research Fund for the Doctoral Program of Higher Education of China (No. 20122304120028).

References

Ahmed SU, Shahjahan M, Murase K (2011) Injecting chaos in feedforward neural networks. Neural Process Lett 34:87–100

Behera L, Kumar S, Patnaik A (2006) On adaptive learning rate that guarantees convergence in feedforward networks. IEEE Trans Neural Netw 17(5):1116–1125

Bertsekas DP, Tsitsiklis JN (2000) Gradient convergence in gradient methods with errors. SIAM J Optim 3:627–642

Charalambous C (1992) Conjugate gradient algorithm for efficient training of artificial neural networks. Inst Electr Eng Proc 139:301–310

Fan QW, Zurada JM, Wu W (2014) Convergence of online gradient method for feedforward neural networks with smoothing L1/2 regularization penalty. Neurocomputing 131:208–216

Fine TL, Mukherjee S (1999) Parameter convergence and learning curves for neural networks. Neural Comput 11:747–769

Guo DQ (2011) Inhibition of rhythmic spiking by colored noise in neural systems. Cogn Neurodyn 5(3):293–300

Hagan MT, Mehnaj MB (1994) Training feedforward networks with Marquardt algorithm. IEEE Trans Neural Netw 5(6):989–993

Haykin S (2008) Neural networks and learning machines. Prentice Hall, New Jersey

Heskes T, Wiegerinck W (1996) A theoretical comparison of batch-mode, on-line, cyclic, and almost-cyclic learning. IEEE Trans Neural Netw 7(4):919–925

Ho KI, Leung CS, Sum JP (2010) Convergence and objective functions of some fault/noise-injection-based online learning algorithms for RBF networks. IEEE Trans Neural Netw 21(6):938–947

Iiguni Y, Sakai H, Tokumaru H (1992) A real-time learning algorithm for a multilayered neural netwok based on extended Kalman filter. IEEE Trans Signal Process 40(4):959–966

Karnin ED (1990) A simple procedure for pruning back-propagation trained neural networks. IEEE Trans Neural Netw 1:239–242

Li Y, Nara S (2008) Novel tracking function of moving target using chaotic dynamics in a recurrent neural network model. Cogn Neurodyn 2(1):39–48

Osowski S, Bojarczak P, Stodolski M (1996) Fast second order learning algorithm for feedforward multilayer neural network and its applications. Neural Netw 9(9):1583–1596

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

Sum JP, Leung CS, Ho KI (2012a) Convergence analyses on on-line weight noise injection-based training algorithms for MLPs. IEEE Trans Neural Netw Learn Syst 23(11):1827–1840

Sum JP, Leung CS, Ho KI (2012b) On-line node fault injection training algorithm for MLP networks: objective function and convergence analysis. IEEE Trans Neural Netw Learn Syst 23(2):211–222

Uwate Y, Nishio Y, Ueta T, Kawabe T, Ikeguchi T (2004) Performance of chaos and burst noises injected to the hopfield NN for quadratic assignment problems. IEICE Trans Fundam E87–A(4):937–943

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

Wu W, Feng G, Li Z, Xu Y (2005) Deterministic convergence of an online gradient method for BP neural networks. IEEE Trans Neural Netw 16:533–540

Wu W, Wang J, Chen MS, Li ZX (2011) Convergence analysis on online gradient method for BP neural networks. Neural Netw 24(1):91–98

Wu Y, Li JJ, Liu SB, Pang JZ, Du MM, Lin P (2013) Noise-induced spatiotemporal patterns in Hodgkin–Huxley neuronal network. Cogn Neurodyn 7(5):431–440

Yoshida H, Kurata S, Li Y, Nara S (2010) Chaotic neural network applied to two-dimensional motion control. Cogn Neurodyn 4(1):69–80

Yu X, Chen QF (2012) Convergence of gradient method with penalty for Ridge Polynomial neural network. Neurocomputing 97:405–409

Zhang NM, Wu W, Zheng GF (2006) Convergence of gradient method with momentum for two-layer feedforward neural networks. IEEE Trans Neural Netw 17(2):522–525

Zhang C, Wu W, Xiong Y (2007) Convergence analysis of batch gradient algorithm for three classes of sigma–pi neural networks. Neural Process Lett 261:77–180

Zhang C, Wu W, Chen XH, Xiong Y (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, Wu W, Yao MC (2012) Boundedness and convergence of batch back-propagation algorithm with penalty for feedforward neural networks. Neurocomputing 89:141–146

Zhang HS, Liu XD, Xu DP, Zhang Y (2014) Convergence analysis of fully complex backpropagation algorithm based on Wirtinger calculus. Cogn Neurodyn 8(3):261–266

Zheng YH, Wang QY, Danca MF (2014) Noise induced complexity: patterns and collective phenomena in a small-world neuronal network. Cogn Neurodyn 8(2):143–149