Convergence of Cyclic and Almost-Cyclic Learning with Momentum for Feedforward Neural Networks

Jian Wang, Jie Yang, and Wei Wu

Abstract— **Two backpropagation algorithms with momentum for feedforward neural networks with a single hidden layer are considered. It is assumed that the training samples are supplied to the network in a cyclic or an almost-cyclic fashion in the learning procedure, i.e., in each training cycle, each sample of the training set is supplied in a fixed or a stochastic order respectively to the network exactly once. A restart strategy for the momentum is adopted such that the momentum coefficient is set to zero at the beginning of each training cycle. Corresponding weak and strong convergence results are then proved, indicating that the gradient of the error function goes to zero and the weight sequence goes to a fixed point, respectively. The convergence conditions on the learning rate, the momentum coefficient, and the activation functions are much relaxed compared with those of the existing results.**

Index Terms— **Almost-cyclic, backpropagation, convergence, cyclic, feedforward neural networks, momentum.**

I. Introduction

LEARNING algorithms play an essential role for feedforward neural networks (NNs). Through learning, the weights of a NN are adapted to meet the requirement of its environment. Backpropagation (BP) method is widely used for training feedforward NNs [1]–[3]. This paper considers two BP algorithms with momentum for feedforward NNs with a single hidden layer.

There are two popular ways of learning with training samples to implement the BP algorithm, batch mode and incremental mode [4]. Corresponding to the standard gradient method, batch mode learning algorithm is completely deterministic but requires additional storage for each weight. On the other hand, incremental mode updates the weights immediately after each sample is fed, and is less demanding on the memory.

There are three incremental learning strategies according to the order in which the samples are applied. The first strategy is online learning (completely stochastic order), i.e., at each learning step, one of the samples is drawn at random from the training set and presented to the network [4]–[7]. The second strategy is almost-cyclic learning (special stochastic order), i.e., the order of sample presentation is continually drawn at random after each training cycle [4], [8]–[10]. The third strategy is cyclic learning (fixed order), that is, in each training cycle, each sample in the training set is supplied in a fixed order, i.e., a particular order of sample presentation is drawn at random before learning starts and then fixed in time [4], [11]–[13].

Some researchers have compared the two basic different training schemes (batch mode and incremental mode) for feedforward NNs [4], [5], [8]. Heskes and Wiegerinck [4] reveal several asymptotic properties of the two schemes and conclude that almost-cyclic learning is a better alternative for batch mode learning than cyclic learning. Wilson [5] explains why batch training is almost always slower than online training (often orders of magnitude slower) especially on large training sets. The main reason is the ability of online training to follow curves in the error surface throughout each cycle, which allows it to safely use a larger learning rate and thus converge with fewer iterations through the training data. Nakama [8] theoretically analyzes the convergence properties of the two schemes applied to quadratic loss functions and shows the exact degrees to which the training set size, the variance of the per-instance gradient, and the learning rate affect the rate of convergence for each scheme.

However, it is well known that a general drawback of gradient-based BP methods is their slow convergence. Many modifications of this learning scheme have been proposed to overcome the difficulty [14], [15]. The BP method with momentum is one of the popular variations. Its idea is to update the weights in the direction, which is a linear combination of the present gradient of the error function and the previous weight updating increment, so as to smooth the weight trajectory and speed up the convergence of the algorithm [16]. It is also sometimes credited with avoiding local minima in the error surface. A recent method of avoiding local minima by convexifying an error criterion is proposed in [17].

There have been some studies on the momentum algorithm in the literature [18]–[24]. Phansalkar and Sastry [18] show that all local minima of the squared error surface are stable points for the BP algorithm with momentum while other equilibrium points are unstable. Hagiwara [19] and Sato [20] show that the momentum coefficient can be derived from a modified cost function, in which the squared errors at the output layer are exponentially weighted in time. They demonstrate a qualitative relationship among the momentum term, the learning rate, and the speed of convergence.

Manuscript received October 14, 2010; revised March 8, 2011; accepted June 11, 2011. Date of publication July 12, 2011; date of current version August 3, 2011. This work was supported in part by the National Natural Science Foundation of China under Grant 10871220 and the China Scholarship Council.

J. Wang is with the School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China. He is also with the Computational Intelligence Laboratory, Department of Electrical and Computer Engineering, University of Louisville, Louisville, KY 40292 USA (e-mail: wangjiannl@mail.dlut.edu.cn).

J. Yang and W. Wu are with the School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China (e-mail: yangjiee@dlut.edu.cn; wuweiw@dlut.edu.cn).

Digital Object Identifier 10.1109/TNN.2011.2159992

Qian [21] shows that the momentum parameter is analogous to the mass of Newtonian particles that move through a viscous medium in a conservative force field. By employing a discrete approximation to the continuous system, Qian also illustrates the conditions for the stability of the algorithm. Torii and Hagan [25] analyze the effect of momentum on steepest descent training for quadratic performance functions. They derive the stability conditions by analyzing the exact momentum equations for the quadratic cost function. In addition, they show a relationship between the momentum coefficient and the speed of convergence of the algorithm. Bhaya [26] points out that the BP with momentum presented in [25] is actually a special case of the more general conjugate gradient method, in which both the learning rate and the momentum coefficient are chosen dynamically in feedback form.

We note that the convergence property for feedforward NN learning is an interesting research topic which offers an effective guarantee in real application. For the gradient-based BP methods without momentum, the existing convergence results focus on online, almost-cyclic, and cyclic learning algorithms. The batch mode learning is essentially a standard gradient descent method. It is easy to see that, if the error criterion is monotonically decreasing, then the convergence of the algorithm is obvious and needs no proof. However, due to the arbitrariness in the presentation order of the training samples, online learning is a completely stochastic process. Thus the convergence results for online learning are mostly asymptotic with a probabilistic nature as the size of training samples goes to infinity [6], [7], [27]–[31]. Deterministic convergence, on the other hand, lies in almost-cyclic and cyclic learning mainly because every sample of the training set is fed exactly once in each training cycle [10]–[13], [32], [33]. It is a bit easier to obtain the convergence for cyclic learning than for almost-cyclic learning. We mention that almost-cyclic learning performs numerically better than cyclic learning since the random property of the training process exists in almost-cyclic learning [4], [9], [10].

Learning rate is an important criterion in the existing convergence analysis. For the training method without momentum, we mention that the special condition of learning rate depends on the different learning fashion. A usual requirement is that the learning rates η_m for online learning satisfy the assumptions $\sum_{m=0}^{\infty}\eta_m = \infty$ and $\sum_{m=0}^{\infty}\eta_m^2 < \infty$ $(\eta_m > 0)$ [30], [31]. In contrast, to obtain the deterministic convergence for cyclic and almost-cyclic learning, authors usually impose certain extra conditions on the learning rate. An additional condition $\lim_{m\to\infty}\eta_m/\eta_{m+1} = 1$ is proposed in [11] to guarantee the convergence for cyclic learning. It is actually a big step forward for the convergence analysis of cyclic learning, compared to the conditions in [12], [13], and [33], which are basically $\eta_m = O(1/m)$. In the existing result for the almost-cyclic learning [10], the condition $\eta_m = O(1/m)$ is still required.

The convergence property of the BP methods with momentum has also been considered by researchers. Bhaya [26] and Torii [25] discuss the convergence of the gradient method with momentum under a restriction that the activation function is linear, which unfortunately is not satisfied by usual activation functions. For the batch learning BP algorithm with momentum, a particular criterion to choose the momentum coefficients term is proposed in [34] and [35] for BP NNs with or without hidden layer, respectively, and the corresponding weak convergence (the gradient of the error function goes to zero) and strong convergence (the weight sequence goes to a fixed point) are proved. The cyclic learning with momentum is considered for feedforward NNs without a hidden layer in [36] and [37], where some tight conditions are required to guarantee the convergence. The learning rates satisfy $0 < \eta_0 \leq 1$ and $1/\eta_{m+1} = 1/\eta_m + N$, where N is a positive constant, and the momentum coefficient $\tau_{m,k}$ for the kth sample at mth learning cycle is chosen as

$$\tau_{m,k} = \begin{cases} \frac{\eta_m^2 \| p^{(m,k)} \|}{\| \Delta \mathbf{w}^{mJ+k-1} \|}, & \text{if } \| \Delta \mathbf{w}^{mJ+k-1} \| \neq 0, \\ 0, & \text{else} \end{cases} \tag{1}$$

where J is the total number of samples, and $p^{(m,k)}$ and $\Delta \mathbf{w}^{mJ+k-1}$ denote the gradient of the error function and the previous increment of the weight. To our best knowledge, there is no convergence results for almost-cyclic learning with momentum.

In this paper, we present a comprehensive study on the weak and strong convergence results for cyclic and almost-cyclic learning with momentum for feedforward NNs with a hidden layer in a quite general framework. Our convergence results are of global nature in that they are valid for arbitrary initial weights. (As a comparison, the above-mentioned result in [18] can be viewed as a local convergence result.) Unlike the corresponding restrictive conditions in [25], [26], [34], and [35], quite simple and general conditions are required in this paper on the learning rates and the momentum coefficients to guarantee the convergence. And these conditions are satisfied by all typical activation functions. In the following paragraphs, we list and explain in detail the main points of the novel contributions of this paper.

1) *The condition on the learning rate for cyclic and almost-cyclic learning with momentum is extended to a more general case:* $\sum_{m=0}^{\infty}\eta_m = \infty$, $\sum_{m=0}^{\infty}\eta_m^2 < \infty$ $(\eta_m > 0)$*, which is identical to those in [6], [7], and [27]–[31] for online learning without momentum.*

We note that the existing convergence results for cyclic and almost-cyclic learning without momentum [10]–[12] are special cases of the momentum methods when the momentum coefficients are set to zero. In a recent convergence result [11] for cyclic learning, an extra condition $\lim_{m\to\infty}\eta_m/\eta_{m+1} = 1$ is required. And for the almost-cyclic learning without momentum [10], a special condition $1/\eta_{m+1} = (1/\eta_m) + l$, $(l > 0)$ on the learning rates is required, which basically means $\eta_m = O(1/m)$. The convergence results in [36] and [37] for cyclic learning with momentum focus on two-layer feedforward NNs, and require $1/\eta_{m+1} = (1/\eta_m) + N$ (N is a positive constant) and $0 < \eta_0 \leq 1$. It is obvious to see that the conditions on the learning rate are much relaxed in this paper than those in [10]–[12], [36], and [37].

2) *Our condition for the momentum coefficients* μ_m *to satisfy* $\sum_{m=0}^{\infty}\mu_m^2 < \infty$ *is more relaxed than those in [36] and [37].*

We note that the (1) on the momentum coefficients is not only closely related to the learning rate but also dependent on the error and the gradient of the error. It is easy to verify that $\sum_{m=0}^{\infty} \tau_{m,k}^2 < \infty$ $(k = 1, \ldots, J)$ is valid. Thus, the (1) is actually a very special case of our above condition.

3) *Our convergence results are valid for both cyclic learning and almost-cyclic learning with momentum.*

We notice that almost-cyclic learning performs numerically better than cyclic learning due to the stochastic nature of the training process [9], [10]. To our best knowledge, the weak and strong convergence results in this paper are novel for almost-cyclic learning with momentum.

4) *We assume that the derivatives g' and f' of the activation functions are locally Lipschitz continuous.*

This condition of ours refines the corresponding conditions in [12], [36], and [37], which demand the boundedness of the second derivative g'', and in [11], which needs g' to be Lipschitz continuous and uniformly bounded on the real number field R. The importance of this condition is that it makes our convergence results apply not only to S-S type NNs (both the hidden and output neurons are with sigmoid activation functions), but also to P-P, P-S, and S-P types NNs, where S and P represent sigmoid and polynomial functions, respectively.

5) *The restrictive assumption on the stationary point set of the error function for the strong convergence in [11], [32], and [37] is relaxed, in that our only requirement on this set is that it does not contain any interior point.*

To obtain the strong convergence result, which means that the weight sequence converges to a fixed point, an additional assumption is introduced in [32], [36], and [37]: the gradient of the error function has finitely many stationary points. A relaxed condition is presented in [11]: the gradient of the error function has at most countably infinitely many stationary points. These conditions are much improved in our case.

The remainder of this paper is organized as follows. In the next section, we formulate mathematically the cyclic and almost-cyclic learning with momentum for feedforward NNs. The main convergence results are presented in Section III, and the rigorous proofs of the main results are provided in Section IV. In Section V, we conclude this paper with some remarks.

II. Cyclic and Almost-Cyclic Learning with Momentum

We consider a feedforward NN with three layers. The numbers of neurons for the input, hidden, and output layers are p, n, and 1, respectively. Suppose that the training sample set is $\{\mathbf{x}^j, O^j\}_{j=0}^{J-1} \subset \mathbb{R}^p \times \mathbb{R}$, where $\mathbf{x}^j$ and O^j are the input and the corresponding ideal output of the jth sample, respectively. Let $\mathbf{V} = \left(v_{i,j}\right)_{n\times p}$ be the weight matrix connecting the input and hidden layers, and we write $\mathbf{v}_i = (v_{i1}, v_{i2}, \ldots, v_{ip})^T$ for $i = 1, 2, \ldots, n$. The weight vector connecting the hidden and output layers is denoted by $\mathbf{u} = (u_1, u_2, \ldots, u_n)^T \in \mathbb{R}^n$. To simplify the presentation, we combine the weight matrix $\mathbf{V}$ with the weight vector $\mathbf{u}$, and write $\mathbf{w} = \left(\mathbf{u}^T, \mathbf{v}_1^T, \ldots, \mathbf{v}_n^T\right)^T \in \mathbb{R}^{n(p+1)}$. Let $g,\ f : \mathbb{R} \to \mathbb{R}$ be given activation functions for the hidden and output layers, respectively. For convenience, we introduce the following vector-valued function:

$$G(\mathbf{z}) = (g(z_1), g(z_2), \ldots, g(z_n))^T \quad \forall\ \mathbf{z} \in \mathbb{R}^n. \tag{2}$$

For any given input $\mathbf{x} \in \mathbb{R}^p$, the output of the hidden neurons is $G(\mathbf{V}\mathbf{x})$, and the final actual output is

$$y = f(\mathbf{u} \cdot G(\mathbf{V}\mathbf{x})). \tag{3}$$

For any fixed weight $\mathbf{w}$, the error of the NNs is defined as

$$\begin{aligned} E(\mathbf{w}) &= \frac{1}{2}\sum_{j=0}^{J-1}(O^j - f(\mathbf{u}\cdot G(\mathbf{V}\mathbf{x}^j)))^2 \\ &= \sum_{j=0}^{J-1} f_j(\mathbf{u}\cdot G(\mathbf{V}\mathbf{x}^j)) \end{aligned} \tag{4}$$

where $f_j(t) = 1/2(O^j - f(t))^2,\ j = 0, 1, \ldots, J-1,\ t \in \mathbb{R}$. The gradients of the error function with respect to $\mathbf{u}$ and $\mathbf{v}_i$ are, respectively, given by

$$\begin{aligned} E_{\mathbf{u}}(\mathbf{w}) &= -\sum_{j=0}^{J-1}\left(O^j - y^j\right) f'(\mathbf{u}\cdot G(\mathbf{V}\mathbf{x}^j))G(\mathbf{V}\mathbf{x}^j) \\ &= \sum_{j=0}^{J-1} f_j'(\mathbf{u}\cdot G(\mathbf{V}\mathbf{x}^j))G(\mathbf{V}\mathbf{x}^j) \end{aligned} \tag{5}$$

$$\begin{aligned} E_{\mathbf{v}_i}(\mathbf{w}) &= -\sum_{j=0}^{J-1}\left(O^j - y^j\right) f'(\mathbf{u}\cdot G(\mathbf{V}\mathbf{x}^j))u_i g'(\mathbf{v}_i\cdot\mathbf{x}^j)\mathbf{x}^j \\ &= \sum_{j=0}^{J-1} f_j'(\mathbf{u}\cdot G(\mathbf{V}\mathbf{x}^j))u_i g'(\mathbf{v}_i\cdot\mathbf{x}^j)\mathbf{x}^j. \end{aligned} \tag{6}$$

Write

$$E_{\mathbf{V}}(\mathbf{w}) = \left(E_{\mathbf{v}_1}(\mathbf{w})^T, E_{\mathbf{v}_2}(\mathbf{w})^T, \ldots, E_{\mathbf{v}_n}(\mathbf{w})^T\right)^T, \tag{7}$$

$$E_{\mathbf{w}}(\mathbf{w}) = \left(E_{\mathbf{u}}(\mathbf{w})^T, E_{\mathbf{V}}(\mathbf{w})^T\right)^T. \tag{8}$$

With cyclic learning with momentum, a particular cycle is drawn at random from the set of all the possible cycles and then kept fixed at all times [4]. The detailed cyclic learning algorithm with momentum is presented as follows. Starting from an arbitrary initial weight $\mathbf{w}^0 = (\mathbf{u}^0, \mathbf{V}^0)$, the network weights are updated iteratively by

$$\begin{cases} \mathbf{u}^{mJ+1} = \mathbf{u}^{mJ} + \eta_m \nabla_0 \mathbf{u}^{mJ}, \quad j = 0 \\ \mathbf{u}^{mJ+j+1} = \mathbf{u}^{mJ+j} + \eta_m \nabla_j \mathbf{u}^{mJ+j} \\ \qquad +\mu_m\left(\mathbf{u}^{mJ+j} - \mathbf{u}^{mJ+j-1}\right), \quad j = 1, \ldots, J-1. \end{cases} \tag{9}$$

$$\begin{cases} \mathbf{v}_i^{mJ+1} = \mathbf{v}_i^{mJ} + \eta_m \nabla_0 \mathbf{v}_i^{mJ}, \quad j = 0 \\ \mathbf{v}_i^{mJ+j+1} = \mathbf{v}_i^{mJ+j} + \eta_m \nabla_j \mathbf{v}_i^{mJ+j} \\ \qquad +\mu_m\left(\mathbf{v}_i^{mJ+j} - \mathbf{v}_i^{mJ+j-1}\right), \quad j = 1, \ldots, J-1 \end{cases} \tag{10}$$

where

$$\begin{aligned}\nabla_k \mathbf{u}^{mJ+j} &= (O^k - y^{mJ+j,k}) f'(\mathbf{u}^{mJ+j} \cdot G^{mJ+j,k}) G^{mJ+j,k} \\ &= -f_k'\left(\mathbf{u}^{mJ+j} \cdot G^{mJ+j,\,k}\right) G^{mJ+j,k}, \qquad (11)\\ \nabla_k \mathbf{v}_i^{mJ+j} &= \left(O^k - y^{mJ+j,k}\right) f'(\mathbf{u}^{mJ+j} \cdot G^{mJ+j,k}) u_i^{mJ+j} \\ &\quad g'(\mathbf{v}_i^{mJ+j} \cdot \mathbf{x}^k)\mathbf{x}^k \\ &= -f_k'\left(\mathbf{u}^{mJ+j} \cdot G^{mJ+j,k}\right) u_i^{mJ+j} \\ &\quad g'\left(\mathbf{v}_i^{mJ+j} \cdot \mathbf{x}^k\right)\mathbf{x}^k, \qquad (12)\\ G^{mJ+j,k} &= G(\mathbf{V}^{mJ+j}\mathbf{x}^k), \qquad (13)\\ y^{mJ+j,k} &= f(\mathbf{u}^{mJ+j} \cdot G^{mJ+j,k}), \qquad (14)\\ m \in \mathbb{N};\ i &= 1, 2, \ldots, n;\ j,\ k = 0, 1, \ldots, J-1.\end{aligned}$$

Here the parameters η_m and μ_m are the learning rate and the momentum coefficient, respectively.

With almost-cyclic learning with momentum, subsequent training cycles are drawn at random. Almost-cyclic learning is online learning with training cycles instead of training patterns, i.e., the training samples are supplied in a stochastic order in each cycle. For the mth training cycle, let $\{\mathbf{x}^{m,1}, \mathbf{x}^{m,2}, \ldots, \mathbf{x}^{m,J}\}$ be a stochastic order of the input vectors $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^J\}$. Similar to the above cyclic learning algorithm, starting from an arbitrary initial weight $\mathbf{w}^0 = (\mathbf{u}^0, \mathbf{V}^0)$, the network weights are updated iteratively by

$$\begin{cases} \mathbf{u}^{mJ+1} = \mathbf{u}^{mJ} + \eta_m \nabla_0^m \mathbf{u}^{mJ}, \quad j = 0, \\ \mathbf{u}^{mJ+j+1} = \mathbf{u}^{mJ+j} + \eta_m \nabla_j^m \mathbf{u}^{mJ+j} \\ \quad +\mu_m \left(\mathbf{u}^{mJ+j} - \mathbf{u}^{mJ+j-1}\right), \quad j = 1, \ldots, J-1 \end{cases} \qquad (15)$$

$$\begin{cases} \mathbf{v}_i^{mJ+1} = \mathbf{v}_i^{mJ} + \eta_m \nabla_0^m \mathbf{v}_i^{mJ}, \quad j = 0 \\ \mathbf{v}_i^{mJ+j+1} = \mathbf{v}_i^{mJ+j} + \eta_m \nabla_j^m \mathbf{v}_i^{mJ+j} \\ \quad +\mu_m \left(\mathbf{v}_i^{mJ+j} - \mathbf{v}_i^{mJ+j-1}\right), \quad j = 1, \ldots, J-1 \end{cases} \qquad (16)$$

where

$$\begin{aligned}&\nabla_k^m \mathbf{u}^{mJ+j} \\ &\quad = (O^k - y^{mJ+j,m,k}) f'(\mathbf{u}^{mJ+j} \cdot G^{mJ+j,m,k}) G^{mJ+j,m,k} \\ &\quad = -f_k'\left(\mathbf{u}^{mJ+j} \cdot G^{mJ+j,m,k}\right) G^{mJ+j,m,k}, \qquad (17)\\ &\nabla_k^m \mathbf{v}_i^{mJ+j} \\ &\quad = \left(O^k - y^{mJ+j,m,k}\right) f'(\mathbf{u}^{mJ+j} \cdot G^{mJ+j,m,k}) u_i^{mJ+j} \\ &\qquad g'(\mathbf{v}_i^{mJ+j} \cdot \mathbf{x}^{m,k})\mathbf{x}^{m,k} \\ &\quad = -f_k'\left(\mathbf{u}^{mJ+j} \cdot G^{mJ+j,m,k}\right) u_i^{mJ+j} \\ &\qquad g'\left(\mathbf{v}_i^{mJ+j} \cdot \mathbf{x}^{m,k}\right)\mathbf{x}^{m,k}, \qquad (18)\\ &\quad G^{mJ+j,m,k} = G(\mathbf{V}^{mJ+j}\mathbf{x}^{m,k}), \qquad (19)\\ &\quad y^{mJ+j,m,k} = f(\mathbf{u}^{mJ+j} \cdot G^{mJ+j,m,k}), \qquad (20)\\ &\quad m \in \mathbb{N};\ i = 1, 2, \ldots, n;\ j,\ k = 0, 1, \ldots, J-1.\end{aligned}$$

Remark: A restart strategy for the momentum is adopted here: the momentum coefficient is set to zero at the beginning of each training cycle. A similar restart strategy has been used in [38] for a conjugate gradient method. We note that this restart strategy makes our convergence analysis much easier, while it does not do any harm to the practical convergence of the training procedure.

III. Main Results

Locally Lipschitz Continuous [39]: Function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be Lipschitz near $\mathbf{x} \in \mathbb{R}^n$ if there exist positive numbers K and ϵ such that we obtain $|f(x_2) - f(x_1)| \le K \|\mathbf{x}_2 - \mathbf{x}_1\|_2$ for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbf{x} + \epsilon b(0, 1)$. If f is Lipschitz near every point of its domain, then it is said to be locally Lipschitz continuous.

For any vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$, we write its Euclidean norm as $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^n x_i^2}$. Let $\Omega_0 = \{\mathbf{w} \in \Omega : E_{\mathbf{w}}(\mathbf{w}) = 0\}$ be the stationary point set of the error function $E(\mathbf{w})$, where $\Omega \subset \mathbb{R}^{n(p+1)}$ is a bounded region satisfying (A4) below. Let $\Omega_{0,s} \subset \mathbb{R}$ be the projection of Ω_0 onto the sth coordinate axis

$$\Omega_{0,s} = \left\{w_s \in \mathbb{R} : \mathbf{w} = (w_1, \ldots, w_s, \ldots, w_{n(p+1)})^T \in \Omega_0\right\} \qquad (21)$$

for $s = 1, 2, \ldots, n(p+1)$. To analyze the convergence of the algorithm, we need the following assumptions:

- (A1) $g'(t)$ and $f'(t)$ are local Lipschitz continuous;
- (A2) $\eta_m > 0$, $\sum_{m=0}^{\infty} \eta_m = \infty$, $\sum_{m=0}^{\infty} \eta_m^2 < \infty$;
- (A3) $\mu_m \ge 0$, $\sum_{m=0}^{\infty} \mu_m^2 < \infty$;
- (A4) there exists a bounded region $\Omega \subset \mathbb{R}^n$ such that $\{\mathbf{w}^m\}_{m=0}^{\infty} \subset \Omega$;
- (A5) $\Omega_{0,s}$ does not contain any interior point for every $s = 1, 2, \ldots, n(p+1)$.

Theorem 3.1: Assume that (A1)–(A4) are valid. Then, starting from an arbitrary initial value $\mathbf{w}^0$, the weight sequence $\{\mathbf{w}^m\}$ defined by (9) and (10) or by (15) and (16) satisfies the following weak convergence:

$$\lim_{m \to \infty} \|E_{\mathbf{w}}\left(\mathbf{w}^m\right)\| = 0. \qquad (22)$$

Moreover, if (A5) is also valid, there holds the strong convergence: There exists $\mathbf{w}^* \in \Omega_0$ such that

$$\lim_{m \to \infty} \mathbf{w}^m = \mathbf{w}^*. \qquad (23)$$

Let us make a few remarks on the convergence result. (A1) allows a broad choice for the activation functions. The assumptions on the activation functions in the existing convergence results [11], [12], [36], [37] are special cases of (A1). We note that, typically, S-S type networks (with sigmoid activation functions for both hidden and output neurons) are used for classification problems, and S-P type networks (with sigmoid hidden neurons and linear or other polynomial output neurons) are used for approximation problems. In this paper, we give a uniform treatment for all the types (S-S, S-P, P-S, and P-P) of BP NNs. As indicated in Contributions 1) and 2), the conditions on the learning rates and the momentum coefficients in this paper [see (A2) and (A3)] are less restrictive than those in [34]–[37] [see (1)]. For the strong convergence, our (A5) on Ω_0 allows it to be a finite set, countably infinite set (such as the set of rational numbers), nowhere dense set (such as Cantor set) or even some uncountable dense set (such as

the set of irrational numbers). The corresponding assumptions in [11], [32], [36], and [37] that the set Ω_0 contains finitely many points or at most countably infinitely many points, respectively, are simple and special cases of (A5) in this paper.

IV. Proofs

For convenience of presentation, we demonstrate in detail the convergence proof for the BP method with cycle learning fashion in Section A. For almost-cycle learning fashion, the proof is similar and introduced in the following Section B.

A. Convergence Analysis for Cyclic learning

In this section, we present five useful lemmas for the convergence analysis.

Lemma 4.1: Let $q(x)$ be a function defined on a bounded closed interval $[a, b]$ such that $q'(x)$ is Lipschitz continuous with Lipschitz constant $K > 0$. Then, $q'(x)$ is differentiable almost everywhere in $[a, b]$ and

$$\left|q''(x)\right| \le K, \ a.e.\ [a, b]. \tag{24}$$

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

$$q(x) \le q(x_0) + q'(x_0)(x - x_0) + C(x - x_0)^2, \forall x_0, x \in [a, b]. \tag{25}$$

Proof: Since $q'(x)$ is Lipschitz continuous on $[a, b]$, $q'(x)$ is absolutely continuous, and the derivative $q''(x)$ exists almost everywhere and is integrable on $[a, b]$. Hence, for almost every $x \in [a, b]$

$$\begin{aligned}\left|q''(x)\right| &= \left|\lim_{h\to 0}\frac{q'(x+h)-q'(x)}{h}\right| \\ &= \lim_{h\to 0}\left|\frac{q'(x+h)-q'(x)}{h}\right| \le K.\end{aligned} \tag{26}$$

Using the integral Taylor expansion, we deduce that

$$\begin{aligned}q(x) &= q(x_0) + q'(x_0)(x - x_0) \\ &\quad + (x - x_0)^2 \int_0^1 (1 - t)q''(x_0 + t(x - x_0))\, dt \\ &\le q(x_0) + q'(x_0)(x - x_0) + (x - x_0)^2 \int_0^1 K(1 - t)\, dt \\ &= q(x_0) + q'(x_0)(x - x_0) + C(x - x_0)^2\end{aligned} \tag{27}$$

where $C = (1/2)K, \ x_0, x \in [a, b]$. ∎

Lemma 4.2: Suppose that the learning rate η_m satisfies (A2) and that the sequence $\{a_m\}$ $(m \in \mathbb{N})$ satisfies $a_m \ge 0$, $\sum_{m=0}^{\infty} \eta_m a_m^\beta < \infty$ and $|a_{m+1} - a_m| \le \mu\eta_m$ for some positive constants β and μ. Then we have

$$\lim_{m\to\infty} a_m = 0. \tag{28}$$

Proof: According to (A2), we know that $\eta_m \to 0$ as $m \to \infty$. We claim that $\lim_{k\to\infty} \inf_{m>k} a_m = 0$. Otherwise, if $a_* \equiv \lim_{k\to\infty} \inf_{m>k} a_m \in (0, \infty]$, then by the definition of inferior limit, there exists an integer $M > k$ such that $a_m \ge (a_*/2) > 0$ for $m \ge M$, which leads to

$$\sum_{m=0}^{\infty} \eta_m a_m^\beta \ge \left(\frac{a_*}{2}\right)^\beta \sum_{m\ge M}^{\infty} \eta_m = \infty. \tag{29}$$

This contradicts $\sum_{m=0}^{\infty} \eta_m a_m^\beta < \infty$ and confirms the claim. Next, we claim that $\lim_{k\to\infty} \sup_{m>k} a_m = 0$. Otherwise, there exists $\delta \in (0, \infty]$ such that $\lim_{k\to\infty} \sup_{m>k} a_m = \delta$. Then, for any $0 < \varepsilon < \delta$, we can choose two subsequences $\{a_{i_k}\}$ and $\{a_{j_k}\}$ of $\{a_m\}$ to satisfy (1)$a_{i_k} \in (0, \varepsilon/4)$, $a_{j_k} \in (\varepsilon, \delta)$; (2) $i_k + 1 < j_k < i_{k+1}$; (3) $a_{i_k+1} \in [\varepsilon/4, \varepsilon/2]$. (This can be done because $\lim_{k\to\infty} \inf_{m>k} a_m = 0$, $\lim_{k\to\infty} \sup_{m>k} a_m = \delta$, and $|a_m - a_{m+1}| \le \mu\eta_m \to 0$ as $m \to \infty$.) For any $i_k < m < j_k$, we have $a_m \in [\varepsilon/4, \varepsilon]$. Thus, we conclude that

$$\begin{aligned}\frac{\varepsilon}{2} &\le \left|a_{j_k} - a_{i_k+1}\right| \le |a_{j_k} - a_{j_k-1}| + \cdots + |a_{i_k+2} - a_{i_k+1}| \\ &\le \mu \sum_{m=i_k+1}^{j_k-1} \eta_m \le \mu \sum_{m=i_k+1}^{j_k} \eta_m.\end{aligned}$$

Therefore, we have for all large enough integers k that

$$\begin{aligned}\sum_{m=i_k}^{\infty} \eta_m a_m^\beta &\ge \sum_{m=i_k+1}^{j_k} \eta_m a_m^\beta \ge \left(\frac{\varepsilon}{4}\right)^\beta \sum_{m=i_k+1}^{j_k} \eta_m \\ &\ge \frac{2}{\mu}\left(\frac{\varepsilon}{4}\right)^{\beta+1}.\end{aligned}$$

But this contradicts $\sum_{m=0}^{\infty} \eta_m a_m^\beta < \infty$ and implies our second claim. Finally, the above two claims together clearly lead to the desired estimate (28). ∎

Lemma 4.3: Let $\{b_m\}$ be a bounded sequence satisfying $\lim_{m\to\infty}(b_{m+1} - b_m) = 0$. Write $\gamma_1 = \lim_{n\to\infty} \inf_{m>n} b_m$, $\gamma_2 = \lim_{n\to\infty} \sup_{m>n} b_m$, and $S = \{a \in \mathbb{R} :$ There exists a subsequence $\{b_{i_k}\}$ of $\{b_m\}$ such that $b_{i_k} \to a$ as $k \to \infty\}$. Then we have

$$S = [\gamma_1, \gamma_2]. \tag{30}$$

Proof: It is obvious that $\gamma_1 \le \gamma_2$ and $S \subset [\gamma_1, \gamma_2]$. If $\gamma_1 = \gamma_2$, then (30) follows simply from $\lim_{m\to\infty} b_m = \gamma_1 = \gamma_2$. Let us consider the case $\gamma_1 < \gamma_2$ and proceed to prove that $S \supset [\gamma_1, \gamma_2]$.

For any $a \in (\gamma_1, \gamma_2)$, there exists $\varepsilon > 0$ such that $(a - \varepsilon, a + \varepsilon) \subset (\gamma_1, \gamma_2)$. Noting $\lim_{m\to\infty}(b_{m+1} - b_m) = 0$, we observe that b_m travels to and from between γ_1 and γ_2 with very small pace for all large enough m. Hence, there must be infinite number of points of the sequence $\{b_m\}$ falling into $(a - \varepsilon, a + \varepsilon)$. This implies $a \in S$ and thus $(\gamma_1, \gamma_2) \subset S$. Furthermore, $(\gamma_1, \gamma_2) \subset S$ immediately leads to $[\gamma_1, \gamma_2] \subset S$. This completes the proof. ∎

Let the sequence $\{\mathbf{w}^{mJ+j}\}$ $(m \in \mathbb{N}, \ j = 0, 1, \ldots, J - 1)$ be generated by (9) and (10). We introduce the following notations:

$$R^{m,\,j} = \eta_m\left(\nabla_j \mathbf{u}^{mJ+j} - \nabla_j \mathbf{u}^{mJ}\right), \tag{31}$$

$$r_i^{m,\,j} = \eta_m\left(\nabla_j \mathbf{v}_i^{mJ+j} - \nabla_j \mathbf{v}_i^{mJ}\right), \tag{32}$$

$$\begin{aligned}d^{m,\,l} &= \mathbf{u}^{mJ+l} - \mathbf{u}^{mJ} \\ &= \eta_m \sum_{j=0}^{l-1} \nabla_j \mathbf{u}^{mJ+j} + \mu_m \sum_{j=1}^{l-1}\left(\mathbf{u}^{mJ+j} - \mathbf{u}^{mJ+j-1}\right) \\ &= \eta_m \sum_{j=0}^{l-1} \nabla_j \mathbf{u}^{mJ} + \sum_{j=0}^{l-1} R^{m,\,j}\end{aligned}$$

$$+\mu_m \sum_{j=1}^{l-1} \left(\mathbf{u}^{mJ+j} - \mathbf{u}^{mJ+j-1}\right), \tag{33}$$

$$\begin{aligned} h_i^{m,\,l} &= \mathbf{v}_i^{mJ+l} - \mathbf{v}_i^{mJ} \\ &= \eta_m \sum_{j=0}^{l-1} \nabla_j \mathbf{v}_i^{mJ+j} + \mu_m \sum_{j=1}^{l-1} \left(\mathbf{v}_i^{mJ+j} - \mathbf{v}_i^{mJ+j-1}\right) \\ &= \eta_m \sum_{j=0}^{l-1} \nabla_j \mathbf{v}_i^{mJ} + \sum_{j=0}^{l-1} r_i^{m,\,j} \\ &\quad +\mu_m \sum_{j=1}^{l-1} \left(\mathbf{v}_i^{mJ+j} - \mathbf{v}_i^{mJ+j-1}\right), \end{aligned} \tag{34}$$

$$\psi^{m,\,l,\,j} = G^{mJ+l,\,j} - G^{mJ,\,j}, \tag{35}$$

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

Let constants C_1, C_2, and C_3 be defined by [see (A3) and (A4)]

$$\begin{aligned} &\max_{0\le j\le J-1} \left\{\|\mathbf{x}^j\|, |O^j|\right\} = C_1, \\ &\sup_{m\in\mathbb{N}} \|\mathbf{w}^m\| = C_2, \quad \sup_{m\in\mathbb{N}} \mu_m = C_3. \end{aligned} \tag{36}$$

By (A1), $f_j'(t)$ also satisfies Lipschitz condition for $j = 0, 1, \ldots, J-1$. Furthermore, $g(t)$, $f(t)$, and $f_j(t)$ are all uniformly continuous on any bounded closed interval.

Lemma 4.4: Let (A1), (A3), and (A4) be valid, and let the sequence $\{\mathbf{w}^{mJ+j}\}$ be generated by (9) and (10). Then, there are constants C_4–C_8 such that

$$\left\|G^{mJ+j,\,k}\right\| \le C_4, \tag{37}$$

$$\|d^{m,\,l}\| \le C_5\eta_m, \quad \|\psi^{m,\,l,\,j}\| \le C_6\eta_m, \tag{38}$$

$$\|R^{m,\,j}\| \le C_7\eta_m^2, \quad \|r_i^{m,\,j}\| \le C_8\eta_m^2 \tag{39}$$

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

Proof: According to (36), we have

$$\left|\mathbf{v}_i^{mJ+j} \cdot \mathbf{x}^k\right| \le \left\|\mathbf{v}_i^{mJ+j}\right\| \left\|\mathbf{x}^k\right\| \le C_1C_2 \triangleq D_1. \tag{40}$$

Thus, there exists a positive constant $C_{4,1}$ such that

$$\max_{|t|\le D_1} |g(t)| = C_{4,1}, \tag{41}$$

$$\left\|G^{mJ+j,\,k}\right\| = \left\|G\left(\mathbf{V}^{mJ+j}\mathbf{x}^k\right)\right\| \le \sqrt{n}C_{4,1} \triangleq C_4. \tag{42}$$

It follows from (36) and (42) that

$$\begin{aligned} \left|\mathbf{u}^{mJ+j} \cdot G^{mJ+j,\,k}\right| &\le \left\|\mathbf{u}^{mJ+j}\right\| \left\|G^{mJ+j,\,k}\right\| \\ &\le C_2C_4 \triangleq D_2. \end{aligned} \tag{43}$$

Then, there is a positive constant $C_{5,1}$ such that

$$\max_{|t|\le D_2} \left|f_j'(t)\right| \le C_{5,1}. \tag{44}$$

Furthermore, a combination of (A1), (9), (11), (37), (43), and (44) gives

$$\begin{aligned} \left\|d^{m,\,l}\right\| &= \left\|\mathbf{u}^{mJ+l} - \mathbf{u}^{mJ}\right\| = \left\|\sum_{j=0}^{l-1} \left(\mathbf{u}^{mJ+j+1} - \mathbf{u}^{mJ+j}\right)\right\| \\ &\le \sum_{j=0}^{l-1} \left\|\left(\mathbf{u}^{mJ+j+1} - \mathbf{u}^{mJ+j}\right)\right\| \\ &\le \sum_{j=0}^{l-1} \left\|-\eta_m \sum_{k=0}^{j} \mu_m^{j-k} f_k' \left(\mathbf{u}^{mJ+k} \cdot G^{mJ+k,k}\right) G^{mJ+k,k}\right\| \\ &\le C_5\eta_m \end{aligned} \tag{45}$$

where $C_5 = JC_4C_{5,1}\sum_{k=0}^{J-1} C_3^{J-k-1}$.

Employing (40), we find that there is a positive constant $C_{6,1}$ such that

$$\max_{|t|\le D_1} \left|g'(t)\right| = C_{6,1}. \tag{46}$$

Moreover, we observe that

$$\begin{aligned} \left\|\psi^{m,\,l,\,j}\right\| &= \left\|G^{mJ+l,\,j} - G^{mJ,\,j}\right\| \\ &\le \max_{1\le i\le n} \left|g'(t_i)\right| \left\|\mathbf{x}^j\right\| \sum_{i=1}^{n} \left\|h_i^{m,\,l}\right\| \\ &\le \max_{1\le i\le n} \left|g'(t_i)\right| \left\|\mathbf{x}^j\right\| \sum_{i=1}^{n} \sum_{k=0}^{l-1} \left\|\mathbf{v}_i^{mJ+k+1} - \mathbf{v}_i^{mJ+k}\right\| \\ &\le C_6\eta_m \end{aligned} \tag{47}$$

where $C_6 = nJC_1^2C_2C_{5,1}C_{6,1}^2\sum_{k=0}^{J-1} C_3^{J-k-1}$, $t_i = \mathbf{v}_i^{mJ} \cdot \mathbf{x}^j + \theta_i(\mathbf{v}_i^{mJ+l} - \mathbf{v}_i^{mJ}) \cdot \mathbf{x}^j$, $\theta_i \in (0, 1)$, and $|t_i| \le C_1C_2$, $(i = 1, 2, \ldots, n)$.

Combining $f_j'(t)$'s Lipschitz continuity, (36) and (37), we have

$$\begin{aligned} &\left|f_j'(\mathbf{u}^{mJ+j} \cdot G^{mJ+j,\,j}) - f_j'(\mathbf{u}^{mJ} \cdot G^{mJ+j,\,j})\right| \\ &\quad \le L\left|\mathbf{u}^{mJ+j} \cdot G^{mJ+j,\,j} - \mathbf{u}^{mJ} \cdot G^{mJ+j,\,j}\right| \\ &\quad \le L\left\|d^{m,\,j}\right\| \left\|G^{mJ+j,\,j}\right\| \le LC_4\left\|d^{m,\,j}\right\|, \end{aligned} \tag{48}$$

$$\begin{aligned} &\left|f_j'(\mathbf{u}^{mJ} \cdot G^{mJ+j,\,j}) - f_j'(\mathbf{u}^{mJ} \cdot G^{mJ,\,j})\right| \\ &\quad \le L\left|\mathbf{u}^{mJ} \cdot G^{mJ+j,\,j} - \mathbf{u}^{mJ} \cdot G^{mJ,\,j}\right| \\ &\quad \le L\left\|\mathbf{u}^{mJ}\right\| \left\|\psi^{m,\,j,\,j}\right\| \le LC_2\left\|\psi^{m,\,j,\,j}\right\| \end{aligned} \tag{49}$$

where $L > 0$ is the Lipschitz constant. By the definition of $R^{m,\,j}$, we see that

$$\begin{aligned} R^{m,\,j} &= \eta_m\left(\nabla_j\mathbf{u}^{mJ+j} - \nabla_j\mathbf{u}^{mJ}\right) \\ &= -\eta_m\big(f_j'(\mathbf{u}^{mJ+j} \cdot G^{mJ+j,\,j})G^{mJ+j,\,j} \\ &\quad - f_j'(\mathbf{u}^{mJ} \cdot G^{mJ,\,j})G^{mJ,\,j}\big) \\ &= -\eta_m\big[f_j'(\mathbf{u}^{mJ+j} \cdot G^{mJ+j,\,j})\psi^{m,\,j,\,j} \\ &\quad + \big(f_j'(\mathbf{u}^{mJ+j} \cdot G^{mJ+j,\,j}) - f_j'(\mathbf{u}^{mJ} \cdot G^{mJ+j,\,j})\big)G^{mJ,\,j} \\ &\quad + \big(f_j'(\mathbf{u}^{mJ} \cdot G^{mJ+j,\,j}) - f_j'(\mathbf{u}^{mJ} \cdot G^{mJ,\,j})\big)G^{mJ,\,j}\big]. \end{aligned} \tag{50}$$

Therefore, it follows from (37), (38), (48), and (49) that

$$\begin{aligned}\left\|R^{m,\,j}\right\| &\le \eta_m\left(LC_4^2\left\|d^{m,\,j}\right\| + (C_{5,1}+LC_2C_4)\left\|\psi^{m,\,j,\,j}\right\|\right)\\ &\le \left(LC_4^2C_5 + \left(C_{5,1}+LC_2C_4\right)C_6\right)\eta_m^2\\ &= C_7\eta_m^2 \end{aligned} \tag{51}$$

where $C_7 = LC_4^2C_5 + (C_{5,1}+LC_2C_4)C_6$.

Similarly, we can show the existence of a constant $C_8 > 0$ such that

$$\left\|r_i^{m,\,j}\right\| \le C_8\eta_m^2. \tag{52}$$

■

The next lemma reveals an almost monotonicity of the error function during the training process.

Lemma 4.5: Let the sequence $\{\mathbf{w}^{mJ+j}\}$ be generated by (9) and (10). Under (A1), (A3), and (A4), there holds

$$\begin{aligned}E\left(\mathbf{w}^{(m+1)J}\right) \le\ & E\left(\mathbf{w}^{mJ}\right) - \eta_m\left\|E_{\mathbf{w}}\left(\mathbf{w}^{mJ}\right)\right\|^2\\ &+ C_9\left(\eta_m^2+\mu_m^2\right)\end{aligned} \tag{53}$$

where $C_9 > 0$ is a constant independent of m, η_m, and μ_m.

Proof: By (A1) and Lemma 4.1, we know that $g''(\mathbf{v}_i^{mJ}\cdot\mathbf{x}^j + t(h_i^{mJ}\cdot\mathbf{x}^j))$ is integrable almost everywhere on [0, 1] and

$$\begin{aligned}&f_j'\left(\mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right)\mathbf{u}^{mJ}\cdot\psi^{m,\,J,\,j}\\ &= f_j'\left(\mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right)\sum_{i=1}^n u_i^{mJ}g'(\mathbf{v}_i^{mJ}\cdot\mathbf{x}^j)h_i^{m,\,J}\cdot\mathbf{x}^j\\ &\quad + f_j'\left(\mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right)\sum_{i=1}^n u_i^{mJ}\left(h_i^{m,\,J}\cdot\mathbf{x}^j\right)^2\\ &\quad\cdot\int_0^1(1-t)g''\left(\mathbf{v}_i^{mJ}\cdot\mathbf{x}^j + t\left(h_i^{m,\,J}\cdot\mathbf{x}^j\right)\right)dt.\end{aligned} \tag{54}$$

By virtue of Lemma 4.1, (11), (12), and (54), there is a constant $C_9 > 0$ such that

$$\begin{aligned}&f_j\left(\mathbf{u}^{(m+1)J}\cdot G^{(m+1)J,\,j}\right)\\ &\le f_j\left(\mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right)\\ &\quad + f_j'\left(\mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right)\left(\mathbf{u}^{(m+1)J}\cdot G^{(m+1)J,\,j} - \mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right)\\ &\quad + C_{10}\left(\mathbf{u}^{(m+1)J}\cdot G^{(m+1)J,\,j} - \mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right)^2\\ &= f_j(\mathbf{u}^{mJ}\cdot G^{mJ,\,j}) + f_j'(\mathbf{u}^{mJ}\cdot G^{mJ,\,j})\\ &\qquad\cdot\left(d^{m,\,J}\cdot G^{mJ,\,j} + \mathbf{u}^{mJ}\cdot\psi^{m,\,J,\,j} + d^{m,\,J}\cdot\psi^{m,\,J,\,j}\right)\\ &\quad + C_{10}\left(\mathbf{u}^{(m+1)J}\cdot G^{(m+1)J,\,j} - \mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right)^2\\ &= f_j(\mathbf{u}^{mJ}\cdot G^{mJ,\,j}) - \nabla_j\mathbf{u}^{mJ}\cdot d^{m,\,J} - \sum_{i=1}^n\left(\nabla_j\mathbf{v}_i^{mJ}\cdot h_i^{m,\,J}\right)\\ &\quad + f_j'\left(\mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right)\sum_{i=1}^n u_i^{mJ}\left(h_i^{m,\,J}\cdot\mathbf{x}^j\right)^2\\ &\qquad\cdot\int_0^1(1-t)g''\left(\mathbf{v}_i^{mJ}\cdot\mathbf{x}^j + t\left(h_i^{m,\,J}\cdot\mathbf{x}^j\right)\right)dt\\ &\quad + f_j'(\mathbf{u}^{mJ}\cdot G^{mJ,\,j})d^{m,\,J}\cdot\psi^{m,\,J,\,j}\\ &\quad + C_{10}\left(\mathbf{u}^{(m+1)J}\cdot G^{(m+1)J,\,j} - \mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right)^2.\end{aligned} \tag{55}$$

Summing (55) from $j = 0$ to $j = J-1$ and noting (4)–(6), (33) and (34), we have

$$\begin{aligned}&E\left(\mathbf{w}^{(m+1)J}\right)\\ &\le E\left(\mathbf{w}^{mJ}\right) - \eta_m\left(\left\|\sum_{j=0}^{J-1}\nabla_j\mathbf{u}^{mJ}\right\|^2 + \sum_{i=1}^n\left\|\sum_{j=0}^{J-1}\nabla_j\mathbf{v}_i^{mJ}\right\|^2\right)\\ &\quad+\delta_m\\ &= E\left(\mathbf{w}^{mJ}\right) - \eta_m\left(\left\|E_{\mathbf{u}}\left(\mathbf{w}^{mJ}\right)\right\|^2 + \sum_{i=1}^n\left\|E_{\mathbf{v}_i}\left(\mathbf{w}^{mJ}\right)\right\|^2\right)\\ &\quad+\delta_m\\ &= E\left(\mathbf{w}^{mJ}\right) - \eta_m\left\|E_{\mathbf{w}}\left(\mathbf{w}^{mJ}\right)\right\|^2 + \delta_m\end{aligned} \tag{56}$$

where

$$\begin{aligned}\delta_m =& -\sum_{j=0}^{J-1}\nabla_j\mathbf{u}^{mJ}\cdot\sum_{j=0}^{J-1}R^{m,\,j}\\ &- \mu_m\sum_{j=0}^{J-1}\nabla_j\mathbf{u}^{mJ}\cdot\sum_{j=1}^{J-1}\left(\mathbf{u}^{mJ+j} - \mathbf{u}^{mJ+j-1}\right)\\ &- \sum_{i=1}^n\left(\sum_{j=0}^{J-1}\nabla_j\mathbf{v}_i^{mJ}\cdot\sum_{j=0}^{J-1}r_i^{m,\,j}\right)\\ &- \mu_m\sum_{i=1}^n\left(\sum_{j=0}^{J-1}\nabla_j\mathbf{v}_i^{mJ}\cdot\sum_{j=1}^{J-1}\left(\mathbf{v}_i^{mJ+j} - \mathbf{v}_i^{mJ+j-1}\right)\right)\\ &+ \sum_{j=0}^{J-1}\sum_{i=1}^n u_i^{mJ}f_j'\left(\mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right)\left(h_i^{m,\,J}\cdot\mathbf{x}^j\right)^2\\ &\qquad\cdot\int_0^1(1-t)g''\left(\mathbf{v}_i^{mJ}\cdot\mathbf{x}^j + t\left(h_i^{m,\,J}\cdot\mathbf{x}^j\right)\right)dt\\ &+ \sum_{j=0}^{J-1}f_j'(\mathbf{u}^{mJ}\cdot G^{mJ,\,j})d^{m,\,J}\cdot\psi^{m,\,J,\,j}\\ &+ C_{10}\sum_{j=0}^{J-1}\left(\mathbf{u}^{(m+1)J}\cdot G^{(m+1)J,\,j} - \mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right)^2.\end{aligned}$$

It now follows from (36) and (37) that

$$\left\|G^{mJ,\,j}\right\| = \left\|G\left(\mathbf{V}^{mJ}\mathbf{x}^j\right)\right\| \le C_4, \tag{57}$$

$$\left|\mathbf{u}^{mJ}\cdot G^{mJ,\,j}\right| \le \left\|\mathbf{u}^{mJ}\right\|\left\|G^{mJ,,j}\right\| \le C_2C_4 = D_2. \tag{58}$$

By (5), (42), (44), and (51), the first term of δ_m can be estimated as follows:

$$\begin{aligned}&\left|-\sum_{j=0}^{J-1}\nabla_j\mathbf{u}^{mJ}\cdot\sum_{j=0}^{J-1}R^{m,\,j}\right|\\ &\qquad\le\left\|E_{\mathbf{u}}\left(\mathbf{w}^{mJ}\right)\right\|\sum_{j=0}^{J-1}\left\|R^{m,\,j}\right\| \le C_{9,1}\eta_m^2\end{aligned} \tag{59}$$

where $C_{9,1} = J^2C_4C_{5,1}C_7$.

Similarly, the second term of δ_m can be obtained as follows:

$$\begin{aligned}&\left|-\mu_m \sum_{j=0}^{J-1} \nabla_j \mathbf{u}^{mJ} \cdot \sum_{j=1}^{J-1} \left(\mathbf{u}^{mJ+j} - \mathbf{u}^{mJ+j-1}\right)\right| \\ &\le \mu_m \left\| E_{\mathbf{u}}\left(\mathbf{w}^{mJ}\right)\right\| \sum_{j=1}^{J-1} \left\|\mathbf{u}^{mJ+j} - \mathbf{u}^{mJ+j-1}\right\| \\ &\le C_{9,2}\left(\eta_m^2 + \mu_m^2\right) \end{aligned} \tag{60}$$

where $C_{9,2} = J^2 C_4^2 C_{5,1}^2 \sum_{k=0}^{J} C_3^{J-k}$.

The estimates for the other terms of δ_m can be obtained with corresponding constants $C_{9,t} > 0$ for $t = 3, \ldots, 7$. Finally, the desired estimate (53) is proved by setting $C_9 = \sum_{t=1}^{7} C_{9,t}$. ■

Now, we are ready to prove the convergence theorem.

Proof to Theorem 3.1: The proof is divided into two parts, dealing with (22) and (23) respectively.

Proof to (22): By (A2), (A3), and Lemma 4.5, we conclude that

$$\sum_{m=0}^{\infty} \eta_m \left\| E_{\mathbf{w}}\left(\mathbf{w}^{mJ}\right)\right\|^2 < \infty, \tag{61}$$

$$\lim_{m\to\infty} \eta_m = 0, \quad \lim_{m\to\infty} \mu_m = 0. \tag{62}$$

By (A1) and (36), it is easy to see that $E_{\mathbf{u}}(\mathbf{w})$ and $E_{\mathbf{v}_i}(\mathbf{w})$ $(i = 1, 2, \ldots, n)$ all satisfy Lipschitz condition. Hence, $E_{\mathbf{w}}(\mathbf{w})$ also satisfies Lipschitz condition with a Lipschitz constant $L' > 0$. By (38), we have

$$\begin{aligned}&\left|\left\| E_{\mathbf{w}}\left(\mathbf{w}^{(m+1)J}\right)\right\| - \left\| E_{\mathbf{w}}\left(\mathbf{w}^{mJ}\right)\right\|\right| \\ &\le \left\| E_{\mathbf{w}}\left(\mathbf{w}^{(m+1)J}\right) - E_{\mathbf{w}}\left(\mathbf{w}^{mJ}\right)\right\| \\ &\le L'\left(\left\| d^{m,J}\right\| + \sum_{i=1}^{n} \left\| h_i^{m,J}\right\|\right) \\ &\le L'\left(C_5 + nJC_1C_2C_{5,1}C_{6,1}\sum_{k=0}^{J} C_3^{J-k}\right)\eta_m \\ &= C_{11}\eta_m \end{aligned} \tag{63}$$

where $C_{11} = L'(C_5 + nJC_1C_2C_{5,1}C_{6,1}\sum_{k=0}^{J} C_3^{J-k})$. Employing (61), (63), and Lemma 4.2, we obtain that

$$\lim_{m\to\infty} \left\| E_{\mathbf{w}}\left(\mathbf{w}^{mJ}\right)\right\| = 0. \tag{64}$$

Similar to (63), there exists a constant $C_{12} > 0$ such that

$$\begin{aligned}&\left\| E_{\mathbf{w}}\left(\mathbf{w}^{mJ+j}\right) - E_{\mathbf{w}}\left(\mathbf{w}^{mJ}\right)\right\| \le C_{12}\eta_m, \\ &j = 0, 1, \ldots, J-1. \end{aligned} \tag{65}$$

It is easy to see that

$$\begin{aligned}&\left\| E_{\mathbf{w}}\left(\mathbf{w}^{mJ+j}\right)\right\| \\ &\le \left\| E_{\mathbf{w}}\left(\mathbf{w}^{mJ+j}\right) - E_{\mathbf{w}}\left(\mathbf{w}^{mJ}\right)\right\| + \left\| E_{\mathbf{w}}\left(\mathbf{w}^{mJ}\right)\right\| \\ &\le C_{12}\eta_m + \left\| E_{\mathbf{w}}\left(\mathbf{w}^{mJ}\right)\right\|. \end{aligned} \tag{66}$$

By (62), (64), and (66), we have

$$\lim_{m\to\infty} \left\| E_{\mathbf{w}}\left(\mathbf{w}^{mJ+j}\right)\right\| = 0, \quad j = 1, 2, \ldots, J-1. \tag{67}$$

This immediately gives

$$\lim_{m\to\infty} \left\| E_{\mathbf{w}}(\mathbf{w}^m)\right\| = 0. \tag{68}$$

■

Proof to (23): According to (A4), the sequence $\{\mathbf{w}^m\}$ $(m \in \mathbb{N})$ has a subsequence $\{\mathbf{w}^{m_k}\}$ $(k \in \mathbb{N})$ that is convergent to, say, $\mathbf{w}^* \in \Omega_0$. It follows from (22) and the continuity of $E_{\mathbf{w}}(\mathbf{w})$ that

$$\begin{aligned}\left\| E_{\mathbf{w}}\left(\mathbf{w}^*\right)\right\| &= \lim_{k\to\infty} \left\| E_{\mathbf{w}}\left(\mathbf{w}^{m_k}\right)\right\| \\ &= \lim_{m\to\infty} \left\| E_{\mathbf{w}}\left(\mathbf{w}^{m}\right)\right\| = 0. \end{aligned} \tag{69}$$

This implies that $\mathbf{w}^*$ is a stationary point of $E(\mathbf{w})$. Hence, $\{\mathbf{w}^m\}$ has at least one accumulation point and every accumulation point must be a stationary point.

Next, by reduction to absurdity, we prove that $\{\mathbf{w}^m\}$ has precisely one accumulation point. Let us assume the contrary that $\{\mathbf{w}^m\}$ has at least two accumulation points $\overline{\mathbf{w}} \neq \widetilde{\mathbf{w}}$. We write $\mathbf{w}^m = (w_1^m, w_2^m, \ldots, w_{n(p+1)}^m)^T$. It is easy to see from (9)–(12) that $\lim_{m\to\infty} \|\mathbf{w}^{m+1} - \mathbf{w}^m\| = 0$ or equivalently, $\lim_{m\to\infty} |w_i^{m+1} - w_i^m| = 0$ for $i = 1, 2, \ldots, n(p+1)$. Without loss of generality, we assume that the first components of $\overline{\mathbf{w}}$ and $\widetilde{\mathbf{w}}$ do not equal to each other, that is, $\bar{w}_1 \neq \tilde{w}_1$. For any real number $\lambda \in (0, 1)$, let $w_1^\lambda = \lambda\bar{w}_1 + (1-\lambda)\tilde{w}_1$. By Lemma 4.3, there exists a subsequence $\{w_1^{m_{k_1}}\}$ of $\{w_1^m\}$ converging to w_1^λ as $k_1 \to \infty$. Due to the boundedness of $\{w_2^{m_{k_1}}\}$, there is a convergent subsequence $\{w_2^{m_{k_2}}\} \subset \{w_2^{m_{k_1}}\}$. We define $w_2^\lambda = \lim_{k_2\to\infty} w_2^{m_{k_2}}$. Repeating this procedure, we end up with decreasing subsequences $\{m_{k_1}\} \supset \{m_{k_2}\} \supset \cdots \supset \{m_{k_{n(p+1)}}\}$ with $w_i^\lambda = \lim_{k_i\to\infty} w_i^{m_{k_i}}$ for each $i = 1, 2, \ldots, n(p+1)$. Write $\mathbf{w}^\lambda = (w_1^\lambda, w_2^\lambda, \ldots, w_{n(p+1)}^\lambda)^T$. Then, we see that $\mathbf{w}^\lambda$ is an accumulation point of $\{\mathbf{w}^m\}$ for any $\lambda \in (0, 1)$. But this means that $\Omega_{0,1}$ has interior points, which contradicts (A5). Thus, $\mathbf{w}^*$ must be a unique accumulation point of $\{\mathbf{w}^m\}_{m=0}^{\infty}$. This completes the proof of the strong convergence. ■

B. Convergence Analysis for Almost-Cyclic Learning

Now, let the sequence $\{\mathbf{w}^{mJ+j}\}$ $(m \in \mathbb{N},\ j = 0, 1, \ldots, J-1)$ be generated by (15) and (16). We introduce the following notations:

$$R^{m,j} = \eta_m\left(\nabla_j^m \mathbf{u}^{mJ+j} - \nabla_j^m \mathbf{u}^{mJ}\right), \tag{70}$$

$$r_i^{m,j} = \eta_m\left(\nabla_j^m \mathbf{v}_i^{mJ+j} - \nabla_j^m \mathbf{v}_i^{mJ}\right), \tag{71}$$

$$\begin{aligned} d^{m,l} &= \mathbf{u}^{mJ+l} - \mathbf{u}^{mJ} \\ &= \eta_m \sum_{j=0}^{l-1} \nabla_j^m \mathbf{u}^{mJ+j} + \mu_m \sum_{j=1}^{l-1}\left(\mathbf{u}^{mJ+j} - \mathbf{u}^{mJ+j-1}\right) \\ &= \eta_m \sum_{j=0}^{l-1} \nabla_j^m \mathbf{u}^{mJ} + \sum_{j=0}^{l-1} R^{m,j} \\ &\quad + \mu_m \sum_{j=1}^{l-1}\left(\mathbf{u}^{mJ+j} - \mathbf{u}^{mJ+j-1}\right), \end{aligned} \tag{72}$$

$$\begin{aligned} h_i^{m,l} &= \mathbf{v}_i^{mJ+l} - \mathbf{v}_i^{mJ} \\ &= \eta_m \sum_{j=0}^{l-1} \nabla_j^m \mathbf{v}_i^{mJ+j} + \mu_m \sum_{j=1}^{l-1} \left(\mathbf{v}_i^{mJ+j} - \mathbf{v}_i^{mJ+j-1} \right) \\ &= \eta_m \sum_{j=0}^{l-1} \nabla_j^m \mathbf{v}_i^{mJ} + \sum_{j=0}^{l-1} r_i^{m,j} \\ &\quad + \mu_m \sum_{j=1}^{l-1} \left(\mathbf{v}_i^{mJ+j} - \mathbf{v}_i^{mJ+j-1} \right), \end{aligned} \tag{73}$$

$$\psi^{m,l,j} = G^{mJ+l,m,j} - G^{mJ,m,j}, \tag{74}$$

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

It is obvious that Lemmas 4.1–4.3 are not influenced by the new definitions. In place of Lemmas 4.4 and 4.5, we now have the following two lemmas.

Lemma 4.6: Let (A1), (A3), and (A4) be valid, and let the sequence $\{\mathbf{w}^{mJ+j}\}$ be generated by (15) and (16). Then, there hold the following estimates with the same constants C_4–C_8 as in Lemma 4.4:

$$\left\| G^{mJ+j,\,m,\,k} \right\| \le C_4, \tag{75}$$

$$\|d^{m,\,l}\| \le C_5 \eta_m, \quad \|\psi^{m,\,l,\,j}\| \le C_6 \eta_m, \tag{76}$$

$$\|R^{m,\,j}\| \le C_7 \eta_m^2, \quad \|r_i^{m,\,j}\| \le C_8 \eta_m^2 \tag{77}$$

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

Proof: According to (36), we have

$$\left| \mathbf{v}_i^{mJ+j} \cdot \mathbf{x}^{m,\,k} \right| \le \left\| \mathbf{v}_i^{mJ+j} \right\| \max_{1 \le k \le J} \left\| \mathbf{x}^k \right\| \le C_1 C_2 \equiv D_1. \tag{78}$$

Thus, there exists a positive constant $C_{4,1}$ such that

$$\max_{|t| \le D_1} |g(t)| = C_{4,1}, \tag{79}$$

$$\left\| G^{mJ+j,\,m,\,k} \right\| = \left\| G\left(\mathbf{V}^{mJ+j} \mathbf{x}^{m,\,k} \right) \right\| \le \sqrt{n} C_{4,1} \equiv C_4. \tag{80}$$

Similarly, (76) and (77) can be proved after adjusting the corresponding superscripts in the proof to Lemma 4.4.

Lemma 4.7: Let the sequence $\{\mathbf{w}^{mJ+j}\}$ be generated by (15) and (16). Under (A1), (A3), and (A4), there holds

$$\begin{aligned} E\left(\mathbf{w}^{(m+1)J} \right) \le\ & E\left(\mathbf{w}^{mJ} \right) - \eta_m \left\| E_{\mathbf{w}}\left(\mathbf{w}^{mJ} \right) \right\|^2 \\ & + C_9 \left(\eta_m^2 + \mu_m^2 \right) \end{aligned} \tag{81}$$

where $C_9 > 0$ is the same constant defined in Lemma 4.5.

Proof: As in the proof to Lemma 4.6, the results are valid as long as the corresponding superscripts be adjusted. The details are left to the interested readers.

Proof of Theorem 3.1 for Almost-Cyclic Learning: The weak and strong convergence results for almost-cyclic learning with momentum can be similarly obtained in terms of Lemmas 4.1–4.3 and Lemmas 4.6–4.7

V. Conclusion

In this paper, the cyclic and almost-cyclic learning algorithms with momentum for three-layer BP neural networks were considered, and a comprehensive study on their weak and strong convergence was carried out. Compared with the existing convergence results, the assumptions to guarantee the convergence are much relaxed, and are valid for more extensive classes of feedforward NNs.

Acknowledgment

The authors would like to thank the reviewers and J. M. Zurada for valuable advice and assistance in the preparation of this paper.

References

[1] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 2001.

[2] D. E. Rumelhart and J. L. McClelland, *Parall Distributed Processing-Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press, 1986.

[3] E. A. de Oliveira and R. C. Alamino, "Performance of the Bayesian online algorithm for the perceptron," *IEEE Trans. Neural Netw.*, vol. 18, no. 3, pp. 902–905, May 2007.

[4] T. Heskes and W. Wiegerinck, "A theoretical comparison of batch-mode, on-line, cyclic, and almost-cyclic learning," *IEEE Trans. Neural Netw.*, vol. 7, no. 4, pp. 919–925, Jul. 1996.

[5] D. R. Wilson and T. R. Martinez, "The general inefficiency of batch training for gradient descent learning," *Neural Netw.*, vol. 16, no. 10, pp. 1429–1451, Dec. 2003.

[6] D. S. Terence, "Optimal unsupervised learning in a single-layer linear feedforward neural network," *Neural Netw.*, vol. 2, no. 6, pp. 459–473, 1989.

[7] W. Finnoff, "Diffusion approximations for the constant learning rate backpropagation algorithm and resistance to local minima," *Neural Comput.*, vol. 6, no. 2, pp. 285–295, Mar. 1994.

[8] T. Nakama, "Theoretical analysis of batch and on-line training for gradient descent learning in neural networks," *Neurocomputing*, vol. 73, nos. 1–3, pp. 151–159, Dec. 2009.

[9] Z. X. Li, W. Wu, and W. Q. Chen, "Prediction of stock market by BP neural networks with technical indexes as input," *J. Math. Res. Explos.*, vol. 23, no. 1, pp. 83–97, 2003.

[10] Z. X. Li, W. Wu, and Y. L. Tian, "Convergence of an online gradient method for feedforward neural networks with stochastic inputs," *J. Comput. Appl. Math.*, vol. 163, no. 1, pp. 165–176, Feb. 2004.

[11] Z.-B. Xu, R. Zhang, and W.-F. Jin, "When does online BP training converge?" *IEEE Trans. Neural Netw.*, vol. 20, no. 10, pp. 1529–1539, Oct. 2009.

[12] W. Wu, G. R. Feng, Z. X. Li, and Y. S. Xu, "Deterministic convergence of an online gradient method for BP neural networks," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 533–540, May 2005.

[13] W. Wu and Y. S. Xu, "Deterministic convergence of an online gradient method for neural networks," *J. Comput. Appl. Math.*, vol. 144, nos. 1–2, pp. 335–347, Jul. 2002.

[14] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation*. Reading, MA: Addison-Wesley, 1991.

[15] S. Becker and Y. Le Cun, "Improving the convergence of back-propagation learning with second-order methods," in *Proc. Connect. Models Summer School*, San Mateo, CA, 1989, pp. 29–37.

[16] M. T. Hagan, H. B. Demuth, and M. H. Beale, *Neural Network Design*. Boston, MA: PWS, 1996.

[17] L. J. Ting-Ho, "Convexification for data fitting," *J. Global Optim.*, vol. 46, no. 2, pp. 307–315, Feb. 2010.

[18] V. V. Phansalkar and P. S. Sastry, "Analysis of the back-propagation algorithm with momentum," *IEEE Trans. Neural Netw.*, vol. 5, no. 3, pp. 505–506, May 1994.

[19] N. O. Attoh-Okine, "Analysis of learning rate and momentum term in backpropagation neural network algorithm trained to predict pavement performance," *Adv. Eng. Softw.*, vol. 30, no. 4, pp. 291–302, Apr. 1999.

[20] A. Sato, "Analytical study of the momentum term in a backpropagation algorithm," in *Proc. Int. Conf. Artif. Neural Netw.*, 1991, pp. 617–622.

[21] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural Netw.*, vol. 12, no. 1, pp. 145–151, Jan. 1999.
[22] X.-H. Yu, G.-A. Chen, and S.-X. Cheng, "Dynamic learning rate optimization of the backpropagation algorithm," *IEEE Trans. Neural Netw.*, vol. 6, no. 3, pp. 669–677, May 1995.
[23] X.-H. Yu and G.-A. Chen, "Efficient backpropagation learning using optimal learning rate and momentum," *Neural Netw.*, vol. 10, no. 3, pp. 517–527, Apr. 1997.
[24] S. V. Kamarthi and S. Pittner, "Accelerating neural network training using weight extrapolations," *Neural Netw.*, vol. 12, no. 9, pp. 1285–1299, Nov. 1999.
[25] M. Torii and M. T. Hagan, "Stability of steepest descent with momentum for quadratic functions," *IEEE Trans. Neural Netw.*, vol. 13, no. 3, pp. 752–756, May 2002.
[26] A. Bhaya and E. Kaszkurewicz, "Steepest descent with momentum for quadratic functions is a version of the conjugate gradient method," *Neural Netw.*, vol. 17, no. 1, pp. 65–71, Jan. 2004.
[27] Y. C. Liang, D. P. Feng, H. P. Lee, S. P. Lim, and K. H. Lee, "Successive approximation training algorithm for feedforward neural networks," *Neurocomputing*, vol. 42, nos. 1–4, pp. 311–322, Jan. 2002.
[28] D. Chakraborty and N. R. Pal, "A novel training scheme for multilayered perceptrons to realize proper generalization and incremental learning," *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 1–14, Jan. 2003.
[29] T. L. Fine and S. Mukherjee, "Parameter convergence and learning curves for neural networks," *Neural Comput.*, vol. 11, no. 3, pp. 747–769, Apr. 1999.
[30] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific, 1996.
[31] V. Tadic and S. Stankovic, "Learning in neural networks by normalized stochastic gradient algorithm: Local convergence," in *Proc. 5th Seminar Neural Netw. Appl. Electron. Eng.*, Belgrade, Yugoslavia, Sep. 2000, pp. 11–17.
[32] W. Wu, H. M. Shao, and D. Qu, "Strong convergence of gradient methods for BP networks training," in *Proc. Int. Conf. Neural Netw. Brains*, Beijing, China, Oct. 2005, pp. 332–334.
[33] W. Wu, G. R. Feng, and X. Li, "Training multilayer perceptrons via minimization of sum of ridge functions," *Adv. Computat. Math.*, vol. 17, no. 4, pp. 331–347, Nov. 2002.
[34] W. Wu, N. M. Zhang, and Z. X. Li, "Convergence of gradient method with momentum for back-propagation neural networks," *J. Comput. Math.*, vol. 26, pp. 613–623, Jul. 2008.
[35] N. M. Zhang, W. Wu, and G. F. Zheng, "Convergence of gradient method with momentum for two-layer feedforward neural networks," *IEEE Trans. Neural Netw.*, vol. 17, no. 2, pp. 522–525, Mar. 2006.
[36] N. M. Zhang, "Deterministic convergence of an online gradient method with momentum," in *Intelligent Computing*, D.-S. Huang, K. Li, and G. Irwin, Eds. Berlin, Germany: Springer-Verlag, 2006, pp. 94–105.
[37] N. M. Zhang, "An online gradient method with momentum for two-layer feedforward neural networks," *Appl. Math. Comput.*, vol. 212, no. 2, pp. 488–498, Jun. 2009.
[38] M. J. D. Powell, "Restart procedures for the conjugate gradient method," *Math. Program.*, vol. 12, no. 1, pp. 241–254, Dec. 1977.
[39] M. Forti, P. Nistri, and M. Quincampoix, "Generalized neural network for nonsmooth nonlinear programming problems," *IEEE Trans. Circuits Syst. I*, vol. 51, no. 9, pp. 1741–1754, Sep. 2004.

Jian Wang received the B.S. degree in computational mathematics from the China University of Petroleum, Dongying, China, in 2002. Since 2006, he has been working toward the Ph.D. degree in computational mathematics at the Dalian University of Technology, Dalian, China.

He was an Instructor with the School of Mathematics and Computational Science, China University of Petroleum, from 2002 to 2006. Currently, he is sponsored by the China Scholarship Council as a Visiting Scholar in the Department of Electrical and Computer Engineering, University of Louisville, Louisville, KY. His current research interests include numerical optimization and neural networks.

Jie Yang received the B.S. degree in computational mathematics from Shanxi University, Taiyuan, China, in 2001, and the Ph.D. degree from the Department of Applied Mathematics, Dalian University of Technology, Dalian, China, in 2006.

She is currently a Lecturer at the School of Mathematical Sciences, Dalian University of Technology. Her current research interests include fuzzy sets and systems, fuzzy neural networks, and spiking neural networks.

Wei Wu received the Bachelor's and Master's degrees from Jilin University, Changchun, China, in 1974 and 1981, respectively, and the Ph.D. degree from Oxford University, Oxford, U.K., in 1987.

He is currently with the School of Mathematical Sciences, Dalian University of Technology, Dalian, China. He has published four books and 90 research papers. His current research interests include learning methods of neural networks.