Global Convergence of Online BP Training with Dynamic Learning Rate

Rui Zhang, Zong-Ben Xu, Guang-Bin Huang, *Senior Member, IEEE*, and Dianhui Wang, *Senior Member, IEEE*

***Abstract*— The online backpropagation (BP) training procedure has been extensively explored in scientific research and engineering applications. One of the main factors affecting the performance of the online BP training is the learning rate. This paper proposes a new dynamic learning rate which is based on the estimate of the minimum error. The global convergence theory of the online BP training procedure with the proposed learning rate is further studied. It is proved that: 1) the error sequence converges to the global minimum error; and 2) the weight sequence converges to a fixed point at which the error function attains its global minimum. The obtained global convergence theory underlies the successful applications of the online BP training procedure. Illustrative examples are provided to support the theoretical analysis.**

***Index Terms*— Backpropagation (BP) neural networks, dynamic learning rate, global convergence analysis, online BP training procedure.**

I. Introduction

THE FEEDFORWARD neural networks trained with the online backpropagation (BP) training procedure have been widely applied in various areas of scientific research and engineering applications [1]–[4]. With the online BP training procedure, all the training examples are sequentially (one-by-one) presented to the learning system and only one training example is learned at each time. The learned example may be randomly or circularly selected from the given training examples, but retained in the training dataset. The online BP training procedure is generally preferred over batch learning in some applications as the network weights are updated immediately after one training example is fed [5], [6]. Therefore, it is worth clarifying the convergence of such online BP training procedure.

Many convergence analyses of various training algorithms for neural networks have been established in these decades. Recently, Ho *et al.* [7] studies the stochastic convergence property of some fault/noise-injection-based online learning for radial basis function networks. Chen *et al.* [8] develop a unified approach for mean-square convergence analysis of ADALINE training with minimum error entropy criterion. In [9], the convergence of the extended Kalman filter-based training for recurrent neural networks has been analyzed. For the online BP training procedure, which is one of the most classical training algorithms for feedforward neural networks, its convergence analysis has been conducted by many researchers and a series of convergence results have been obtained. The probabilistic convergence properties of the online BP training procedure have been studied in [10]–[16], followed by a series of deterministic convergence analyses [17]–[27]. The neural networks discussed in [18], [19], [21], [23], [24], and [27] are implemented without hidden layers, and hence, are of very special form. Although [17], [20], [22] further studied the feedforward networks with hidden layers, the obtained results have only concluded the convergence of the gradient sequence of the error functions but not justified the convergence of the weight sequence itself. Recently, Xu *et al.* [25] revealed the convergence of the weight sequence (named the strong convergence) with dismissing all the posterior assumptions on the setting of the learning rate. Meanwhile, the convergence of the gradient sequence of the error functions (named the weak convergence) has also been proved in [25] under much relaxed conditions than those supposed in [17] and [22].

It should be noted that all the research studies referred to above have focused on the online BP training procedure with a constant learning rate or a diminishing learning rate, that is, the learning rate η_m satisfies either $\eta_m = \alpha$ where α is a constant ([11], [13]), or $\eta_m \to 0$ as $m \to \infty$ with the constraint $\sum_{m=0}^{\infty} \eta_m = \infty$ ([28], such as $\eta_m = 1/m$). One of the main factors affecting the performance of the online BP training procedure in applications is the selected learning rate. Although the constant learning rate is simple and easy to be implemented, it cannot guarantee the convergence of the online BP training procedure in most cases. Therefore, most convergence analyses of the online BP training procedure have been conducted by adopting the diminishing learning rate. A number of convergence results have been established in terms of the nature of the diminishing learning rate. Nevertheless, the diminishing learning rate is generally set as a time-varying

Manuscript received October 9, 2010; revised September 8, 2011; accepted October 29, 2011. Date of publication December 24, 2011; date of current version February 8, 2012. This work was supported in part by the grant from Academic Research Fund Tier 1 of Ministry of Education under Project RG 22/08 (M52040128), Singapore, and the National Basic Research Program of China (973 Program) under Grant 2007CB31002, and the National Natural Science Foundation of China, under Grant 61075050 and Grant 61075054.

R. Zhang is with the School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore. He is also with the Department of Mathematics, Northwest University, Xi'an 710069, China (e-mail: rzhang@nwu.edu.cn).

Z.-B. Xu is with the Institute for Information and System Science and MOE Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: zbxu@mail.xjtu.edu.cn).

G.-B. Huang is with the School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore (e-mail: egbhuang@ntu.edu.sg).

D. H. Wang is with the Department of Computer Science and Computer Engineering, La Trobe University, VIC 3086, Australia (e-mail: dh.wang@latrobe.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2011.2178315

sequence with slow ultimate convergence, i.e., η_m decreases to zero only with the increase of the iterations, but without being related to the error achieved at each iteration at all, which is the optimized objective of the online BP training. From another point of view, although there has been proposed other heuristic method to design an adaptive learning rate based on the evolution of the error [29], the convergence of the online BP training procedure has not been mentioned at all.

In this paper, we will first propose a new scheme to set a dynamic learning rate, by which η_m is adjusted dynamically on the basis of the estimate of the minimum error. The proposed dynamic learning rate is more reasonable and reliable in applications as it utilizes the error information during the training procedure. It can be considered as a key generalization that allows a wider class of learning rates to apply for the online BP training procedure. We will then study the global convergence property of the online BP training procedure with the proposed dynamic learning rate. It will be verified that: 1) the error sequence converges to the global minimum error, which obviously implies the weak convergence result shown in [25]; and 2) the weight sequence converges to a fixed point at which the error function attains its global minimum, which sharpens the strong convergence result in [25] under weaker conditions.

The remainder of this paper is organized as follows. In Section II, we formulate mathematically the online BP training procedure with the dynamic learning rate. We present the main results in Section III. The rigorous proofs of the main results and some useful lemmas are presented in Section IV. Then we show the performance evaluation in Section V and conclude this paper in Section VI.

II. On-Line BP Training with Dynamic Step-Size Rule

Without loss of generality, we begin with an introduction of a single-hidden layer feedforward network with p inputs, n hidden neurons, and 1 output neuron. Denote by $V = (v_{ij})_{n\times p} = (v_1, v_2, \ldots, v_n)^T$ the weight matrix connecting the input layer and the hidden layer where $v_i = (v_{i1}, v_{i2}, \ldots, v_{ip})$ $(i = 1, 2, \ldots, n)$, and by $\omega = (\omega_1, \omega_2, \ldots, \omega_n)^T$ the weight vector connecting the hidden layer and the output layer. All the hidden and output neurons use the same continuously differentiable activation function, which is denoted by $g : R \to R$. Define the following vector-valued function:

$$G(s) = (g(s_1), g(s_2), \ldots, g(s_n))^T \qquad \forall s \in R^n.$$

For any input $x \in R^p$, the output of the hidden layer is $G(Vx - \theta)$ where $\theta = (\theta_1, \theta_2, \ldots, \theta_n)^T$ is the threshold in the hidden layer. Let $\tilde{V} = (V, \theta) \in R^{n\times(p+1)}$ and $\tilde{x} = (x^T, -1)^T \in R^{p+1}$, then we have $G(Vx - \theta) = G(\tilde{V}\tilde{x})$. Therefore, for the sake of simplicity, we can consider that $\theta = 0$. In the same way, the output of the entire network can be written as

$$\begin{aligned} y &= g(\omega \cdot G(Vx)) \\ &= g(\omega_1 g(v_1 x) + \omega_2 g(v_2 x) + \cdots + \omega_n g(v_n x)) \end{aligned}$$

where $\omega \cdot G(Vx)$ denotes the inner product of ω and $G(Vx)$.

Let $D_J = \{(x^j, t^j)\}_{j=0}^{J-1} \subset R^p \times R$ be the given training example set. For any $\omega \in R^n$ and $V \in R^{n\times p}$, the error of the neural network is defined by

$$E(\omega, V) = \sum_{j=0}^{J-1} E_j(\omega, V)$$

where

$$\begin{aligned} E_j(\omega, V) &= \frac{1}{2}\left(t^j - y^j\right)^2 \\ &= \frac{1}{2}\left(t^j - g\left(\omega \cdot G\left(Vx^j\right)\right)\right)^2, j = 0, 1, \ldots, J-1 \end{aligned}$$

which are called the individual error functions.

The objective of training neural networks is to search a set of optimal weights between neurons so as to generate outputs y^j as close as possible to the targets $t^j (j = 0, 1, \ldots, J-1)$, which is equivalent to minimizing the error function $E(\omega, V)$. The BP training scheme is an approach to estimating the optimal weights through applying the gradient descent method, combined with the BP scheme of computation for gradient of the error function [3].

Denoted by $\nabla E_{j,\omega}(\omega, V)$ and $\nabla E_{j,V}(\omega, V)$ the gradients of each individual error function $E_j(\omega, V)$ with respect to ω and V, respectively. We have

$$\begin{aligned} \nabla E_{j,\omega}(\omega, V) &:= \left(\frac{\partial E_j(\omega, V)}{\partial \omega_1}, \ldots, \frac{\partial E_j(\omega, V)}{\partial \omega_n}\right)^{\mathsf{T}} \\ &= -(t^j - y^j) g'(\omega \cdot G(Vx^j)) G(Vx^j) \\ \nabla E_{j,V}(\omega, V) &:= \left(\nabla E_{j,v_1}(\omega, V), \ldots, \nabla E_{j,v_n}(\omega, V)\right)^{\mathsf{T}} \\ \nabla E_{j,v_i}(\omega, V) &:= \left(\frac{\partial E_j(\omega, V)}{\partial v_{i1}}, \ldots, \frac{\partial E_j(\omega, V)}{\partial v_{ip}}\right) \\ &= -(t^j - y^j) g'(\omega \cdot G(Vx^j)) g'(v_i x^j) \omega_i (x^j)^T \\ &\quad i = 1, \ldots, n, \ j = 0, \ldots, J-1. \end{aligned}$$

With the above notations and equations, the online BP training procedure can be formulated as the following iteration (see [25])

$$\begin{cases} \omega^{mJ+j+1} = \omega^{mJ+j} - \eta_m \nabla E_{j,\omega}(\omega^{mJ+j}, V^{mJ+j}) \\ V^{mJ+j+1} = V^{mJ+j} - \eta_m \nabla E_{j,V}(\omega^{mJ+j}, V^{mJ+j}) \\ \qquad j = 0, 1, \ldots, J-1, \ m = 0, 1, \ldots \end{cases}$$

Denoted by (ω, V) the matrix $[\omega : V]$. The above iteration can be equivalently expressed in the matrix form by

$$\begin{aligned} &(\omega^{mJ+j+1}, V^{mJ+j+1}) \\ &= (\omega^{mJ+j}, V^{mJ+j}) - \eta_m \nabla E_j(\omega^{mJ+j}, V^{mJ+j}) \\ &\quad j = 0, 1, \ldots, J-1, m = 0, 1, \ldots \end{aligned} \tag{1}$$

where $\nabla E_j(\omega, V) = (\nabla E_{j,\omega}(\omega, V), \nabla E_{j,V}(\omega, V))$ is the gradient of the individual error function $E_j(\omega, V)$ with respect to (ω, V). In the following, we use the Frobenius norm of matrix (ω, V) which is defined by $\|(\omega, V)\| = (\|\omega\|^2 + \|V\|_F^2)^{1/2}$ with $\|V\|_F^2 = \sum_{i=1}^n \|v_i\|^2 = \sum_{i=1}^n \sum_{j=1}^p v_{ij}^2$ and $\|\omega\| = \sum_{i=1}^n \|\omega_i\|^2$.

In the procedure (1), the online gradient scheme is applied. Specifically, at each iteration, the weight sequence $\{(\omega^{mJ+j}, V^{mJ+j})\}$ is changed incrementally, through a

sequence of J steps. Each step is a gradient iteration for an individual error function $E_j(\omega, V)$ associated with a single training example, and there is one step per individual error function. Thus, an iteration can be viewed as a cycle of J subiterations. In this paper, we only consider the situation of choosing the training examples in a fixed order, that is, from x^0 to x^{J-1} one by one sequentially. Here, η_m is the learning rate of the procedure, which is updated after each cycle of J subiterations. It has been pointed out that a key issue concerning the learning rate η_m is often crucial for the success of the gradient-based algorithms [28], [30]. Therefore, designing a more reasonable and appropriate learning rate, which ensures the convergence of the online BP training procedure meanwhile, is our main concern in this paper.

We first formulate an assumption before introducing the new dynamic learning rate.

Assumption 1: $E(\omega^k, V^k) \to \infty$ as $\|(\omega^k, V^k)\| \to \infty$.

As the error function $E(\omega, V)$ is nonnegative, the global minimum value of $E(\omega, V)$ must exist and be finite. Throughout this paper, we use the notation $E^* = \min E(\omega, V)$ and $\Omega^* = \{(\omega, V) : E(\omega, V) = E^*\}$ to denote the global minimum error and the set of the optimal solutions to $\min E(\omega, V)$. Then Ω^* must be bounded under Assumption 1. In this paper, the dynamic learning rate is formulated as

$$\eta_m = \beta_m \cdot \frac{E(\omega^{mJ}, V^{mJ}) - E_m^{lev}}{C^2} \qquad 0 < \underline{\beta} \le \beta_m \le \overline{\beta} < 2 \tag{2}$$

where C is a positive scalar and E_m^{lev} is an estimate of E^*, which is improved successively on the basis of the function values $E(\omega^{mJ}, V^{mJ})$ observed. E_m^{lev} is given by

$$E_m^{lev} = \min_{\substack{0 \le i \le m-1 \\ 0 \le j \le J-1}} E\left(\omega^{iJ+j+1}, V^{iJ+j+1}\right) - \delta_m \tag{3}$$

where δ_m is an adjustable parameter satisfying

$$\delta_m \to 0^+ \text{ as } m \to \infty \quad \text{and} \quad \sum_{m=0}^{\infty} \delta_m^2 = \infty. \tag{4}$$

It should be noted that the parameter δ_m is actually a key factor to control the decreasing rate of the error sequence $\{E(\omega^{mJ+j}, V^{mJ+j})\}$. In what follows, we explain in detail where the two conditions "$\delta_m \to 0^+$ as $m \to \infty$" and "$\sum_{m=0}^{\infty} \delta_m^2 = \infty$" in (4) come from. Since E_m^{lev} is defined as an estimate of the global minimum error E^*, in general, E_m^{lev} should satisfy

$$E_m^{lev} \to E^* \quad \text{as} \quad m \to \infty \tag{5}$$

and

$$E_m^{lev} \le \min_{\substack{0 \le i \le m-1 \\ 0 \le j \le J-1}} E(\omega^{iJ+j+1}, V^{iJ+j+1}) \quad \text{for all } m = 0, 1, \ldots \tag{6}$$

Then, comparing (3) with (5), we obviously have $\delta_m \to 0$. And comparing (3) with (6), we know that δ_m should be positive. This is the reason why the first condition "$\delta_m \to 0^+$ as $m \to \infty$" is required. On the other hand, δ_m should not decrease too rapidly, otherwise, δ_m will become quite small as m increases, which may have very little effect on the decrease of the error sequence $\{E(\omega^{mJ+j}, V^{mJ+j})\}$. Thus, in order to avoid this situation, the second condition "$\sum_{m=0}^{\infty} \delta_m^2 = \infty$" is supposed.

Remark 1: The proposed dynamic learning rate (2)–(4) is originated from the dynamic step-size rule introduced in [28], which was formulated by

$$\alpha_k = \gamma_k \frac{f(x_k) - f_k^{lev}}{C^2}, \qquad 0 < \underline{\gamma} \le \gamma_k \le \overline{\gamma} < 2$$

where $f(x) = \sum_{i=1}^{m} f_i(x)$ and f_k^{lev} is an estimate of the optimal function value $f^* = \inf f(x_k)$ for the optimization problem $\min f(x)$

$$f_k^{lev} = \min_{0 \le j \le k} f(x_j) - \delta_k.$$

In [28], two distinct procedures for adjusting δ_k were proposed. The first adjustment procedure [see (2.11) in [28]] is simple but is guaranteed to yield only a δ-optimal objective function value with δ positive and arbitrary small (that is, $\inf_{k \ge 0} f(x_k) \le f^* + \delta$ when $f^* > -\infty$), while the second adjustment procedure (see the path-based incremental target level algorithm in [28]) is more complex but is guaranteed to yield the optimal value f^* in the sense that $\inf_{k \ge 0} f(x_k) = f^*$.

Remark 2: In this paper, for the network training problem $\min E(\omega, V) = \sum_{j=0}^{J-1} E_j(\omega, V)$, we present a new adjustment formulation (4) on δ_m, with which not only the convergence of the error sequence $\{E(\omega^{mJ+j}, V^{mJ+j})\}$ will be conducted, but also the weight sequence $\{(\omega^{mJ+j}, V^{mJ+j})\}$ itself will be settled down. Although the topic in this paper focuses on the online BP training procedure of neural networks, which is a special case of particular interest to us in applications, the convergence results we obtained can be easily confirmed to be true for the optimization problems with the general form $\min f(x) = \sum_{i=1}^{m} f_i(x)$ when the objective function $f(x)$ is given by a summation of a finite number of functions $f_i(x) (i = 1, \ldots, m)$.

III. Global Convergence Results

This section summarizes the main results we have obtained for global convergence of the online BP training procedure (1) with dynamic learning rate (2)–(4) under Assumption 1 and the following Assumptions 2 and 3. The proofs of the results are, however, postponed to the next section so as to make the presentation more readable.

Assumption 2: 1) Each individual error function $E_j(\omega, V)$ is convex

$$E_j(\bar{\omega}, \bar{V}) - E_j(\omega, V) \ge \left\langle \nabla E_j(\omega, V), (\bar{\omega}, \bar{V}) - (\omega, V) \right\rangle$$

or equivalently

$$E_j(\omega, V) - E_j(\bar{\omega}, \bar{V}) \le \left\langle \nabla E_j(\omega, V), (\omega, V) - (\bar{\omega}, \bar{V}) \right\rangle$$

for any $(\omega, V), (\bar{\omega}, \bar{V}) \in R^n \times R^{n \times p}$ and $j = 0, 1, \ldots, J-1$.

2) The gradients of each individual error function is bounded, that is, there exist two positive scalars $C_{j,\omega} > 0$ and $C_{j,V} > 0$ such that $\|\nabla E_{j,\omega}(\omega, V)\| \le C_{j,\omega}$ and $\|\nabla E_{j,V}(\omega, V)\| \le C_{j,V}$.

Assumption 3: The weight sequence $\{(\omega^{mJ}, V^{mJ})\}$ has at most countably infinite number of limit points.

As mentioned in Section II, with the online BP training procedure (1), the weight sequence $\{(\omega^{mJ+j},\ V^{mJ+j})\}$ is updated incrementally at each cycle $(m = 0, 1, \ldots)$ through J $(j = 0, 1, \ldots, J-1)$ subiterations. In the subsequent convergence analysis, we first clarify the convergence results for the sequence $\{(\omega^{mJ}, V^{mJ})\}$, which is composed of the elements generated at the beginning of each cycle. Then we extend the obtained results to the general case of the whole weight sequence $\{(\omega^{mJ+j}, V^{mJ+j})\}$.

Theorem 1: Let the sequence $\{(\omega^{mJ}, V^{mJ})\}$ be generated by the online BP training procedure (1) with dynamic learning rate (2)–(4). If the set Ω^* contains finite points $(\omega_1^*, V_1^*), (\omega_2^*, V_2^*), \ldots, (\omega_k^*, V_k^*)$, then under Assumptions 1 and 2, there must exists $(\omega_i^*, V_i^*) \in \Omega^* (i \in \{1, 2, \ldots, k\})$ such that

$$\lim_{m\to\infty} (\omega^{mJ}, V^{mJ}) = (\omega_i^*, V_i^*).$$

Theorem 1 shows that the weight sequence $\{(\omega^{mJ}, V^{mJ})\}$ converges to some minimum point under the assumption that the set Ω^* only contains finite points. We next extend this result to the general case where there is no restriction on Ω^*. We need to first introduce how is a sequence called to be convergent to a set.

Definition 1 [31]: Let $\{x_n\}$ be a sequence and X be a set. If for any positive number $\varepsilon > 0$, there is an integer N such that $\rho(x_n, X) < \varepsilon$ for all $n \geq N$, then the sequence $\{x_n\}$ is said to be convergent to the set X, denoted by $\lim_{n\to\infty} \rho(x_n, X) = 0$. Here, $\rho(x_n, X) = \inf\{\rho(x_n, x) : x \in X\}$ denotes the distance between $\{x_n\}$ and X.

Theorem 2: Let the sequence $\{(\omega^{mJ}, V^{mJ})\}$ be generated by the online BP training procedure (1) with dynamic learning rate (2)–(4). Then under Assumptions 1 and 2, we have

$$\lim_{m\to\infty} \rho((\omega^{mJ}, V^{mJ}), \Omega^*) = 0.$$

Theorem 2 reveals that the weight sequence $\{(\omega^{mJ}, V^{mJ})\}$ converges to the set Ω^*. In the special case where Ω^* is finite, $\lim_{m\to\infty} \rho((\omega^{mJ}, V^{mJ}), \Omega^*) = 0$ is obviously equivalent to the fact that (ω^{mJ}, V^{mJ}) converges to some minimum point in Ω^*, which is consistent with Theorem 1. Therefore, Theorem 1 actually states a special case of Theorem 2.

Suppose that there is a bounded set B such that $(\omega^{mJ}, V^{mJ}) \in B$ for all m [same as Assumption (A2) in [25]] instead of Assumption 1, then Theorem 2 still holds by only using the compact set $\Omega^* \cap B$ rather than Ω^* in the proof. Furthermore, with Theorem 2, we can easily derive the following global convergence result for the error sequence $\{E(\omega^{mJ}, V^{mJ})\}$.

Theorem 3: Under Assumptions 1 and 2, the online BP training procedure (1) with dynamic learning rate (2)–(4) is globally weak convergent

$$\lim_{m\to\infty} E(\omega^{mJ}, V^{mJ}) = E^*.$$

Remark 3: Combining with (2)–(4), we have $\eta_m \geq \beta_m \cdot (\delta_m/C^2) \geq (\underline{\beta}/C^2)\delta_m$ and consequently

$$\sum_{m=0}^{\infty} \eta_m^2 \geq \frac{\underline{\beta}^2}{C^4} \sum_{m=0}^{\infty} \delta_m^2 = \infty$$

as well as

$$\sum_{m=0}^{\infty} \eta_m \geq \frac{\underline{\beta}}{C^2} \sum_{m=0}^{\infty} \delta_m = \infty$$

due to $\sum_{m=0}^{\infty} \delta_m = \infty$ in view that $\delta_m \to 0^+$ and $\sum_{m=0}^{\infty} \delta_m^2 = \infty$. Furthermore, we can also derive that

$$\eta_m = \beta_m \cdot \frac{E(\omega^{mJ}, V^{mJ}) - E_m^{lev}}{C^2} \to 0 \quad \text{as} \quad m \to \infty.$$

Since $E(\omega^{mJ}, V^{mJ}) - E_m^{lev} \to 0$ which follows from the fact that $\lim_{m\to\infty} E(\omega^{mJ}, V^{mJ}) = E^*$. Therefore, the dynamic learning rate (2)–(4) presented in this paper can be viewed as an extension of the diminishing learning rate (say, $\sum_{m=0}^{\infty} \eta_m = \infty$ and $\sum_{m=0}^{\infty} \eta_m^2 < \infty$) adopted in [25]. It is also worth mentioning that the condition $\sum_{m=0}^{\infty} \eta_m^2 = \infty$ rather than $\sum_{m=0}^{\infty} \eta_m^2 < \infty$, to some extent, can avoid slow ultimate convergence from which the diminishing learning rate always suffers. This is another contribution of the proposed dynamic learning rate in this paper.

As a consequence of all the above obtained results, we can prove the final convergence theorem claiming that the weight sequence $\{(\omega^{mJ}, V^{mJ})\}$ converge to some minimum point in Ω^* under a mild condition.

Theorem 4: Under Assumptions 1–3, the online BP training procedure (1) with dynamic learning rate (2)–(4) is globally strong convergent, that is, there must exist a fixed value $(\omega^*, V^*) \in \Omega^*$ such that

$$\lim_{m\to\infty} (\omega^{mJ}, V^{mJ}) = (\omega^*, V^*).$$

Remark 4: According to the procedure (1), Assumption 2 2) and the fact that $\eta_m \to 0$, we can obviously deduce that

$$\begin{aligned}
&\lim_{m\to\infty} \left(\omega^{mJ+j}, V^{mJ+j}\right) \\
&= \lim_{m\to\infty} \Big[\left(\omega^{mJ+j-1}, V^{mJ+j-1}\right) \\
&\qquad -\eta_m \nabla E_{j-1} \left(\omega^{mJ+j-1}, V^{mJ+j-1}\right) \Big] \\
&= \lim_{m\to\infty} \left(\omega^{mJ+j-1}, V^{mJ+j-1}\right)
\end{aligned}$$

for any $j = 1, 2, \ldots, J-1$. On the basis of this relationship, the convergence of the whole weight sequence $\{(\omega^{mJ+j}, V^{mJ+j}),\ j = 0, 1, \ldots, J-1,\ m = 0, 1, \ldots\}$ can be surely guaranteed as long as the weight sequence $\{(\omega^{mJ}, V^{mJ}),\ m = 0, 1, \ldots\}$ converges. Therefore, all the convergence results presented in Theorems 1–4 must still hold for the whole weight sequence $\{(\omega^{mJ+j}, V^{mJ+j})\}$. Moreover, as we explained in the proceeding context, it can be concluded that Theorems 1–4 also hold if we use the assumption that the sequence $\{(\omega^{mJ+j}, V^{mJ+j})\}$ is bounded [Assumption (A2) in [25]] instead of Assumption 1 in this paper.

Remark 5: In [25], the weak convergence result that the gradient sequence of the error functions $\{\nabla E(\omega^{mJ+j}, V^{mJ+j})\}$ generated by the online BP training procedure with diminishing learning rate goes to zero as $m \to \infty$, has been proved. In this paper, however, we verify that the error sequence $\{E(\omega^{mJ+j}, V^{mJ+j})\}$ generated by the online BP training procedure with dynamic learning rate converges to the global

minimum value E^* as $m \to \infty$, which definitely implies that $\nabla E(\omega^{mJ+j}, V^{mJ+j}) \to 0$ as $m \to \infty$ due to the continuity of the gradient. This fact reveals that a stronger weak convergence result is presented in Theorem 3 in this paper together with wider selection of the learning rate for the online BP training procedure, compared with Theorem 1 in [25].

Remark 6: Denoted by $\Omega(E) = \{(\omega, V) : \nabla E(\omega, V) = 0\}$ the set of the stationary points of $E(\omega, V)$ and by W the set of the limit points of the weight sequence $\{(\omega^{mJ}, V^{mJ})\}$. In [25], the strong convergence result that the weight sequence $\{(\omega^{mJ+j}, V^{mJ+j})\}$ converges to a fixed point $(\omega^*, V^*) \in \Omega(E)$ has been shown in Theorem B under the assumption that $\Omega(E)$ is at most infinitely countable. In this paper, we prove that the weight sequence $\{(\omega^{mJ+j}, V^{mJ+j})\}$ converges to a fixed point $(\omega^*, V^*) \in \Omega^*$ in Theorem 4 supposing that W is at most infinitely countable. It is well known that $\nabla E(\omega, V) = 0$ is a necessary condition of (ω, V) being the minimum point of $E(\omega, V)$. Therefore, the strong convergence shown in [25] can only guarantee that the weight sequence $\{(\omega^{mJ+j}, V^{mJ+j})\}$ converges to a local minimum point. However, in this paper, not only the global strong convergence of $\{(\omega^{mJ+j}, V^{mJ+j})\}$ obtained, but also this result is proved under a relaxed condition that "W is at most infinitely countable" rather than the condition that "$\Omega(E)$ is at most infinitely countable."

IV. Proof of the Main Results

The proofs of all the theorems presented in the last section are given in this section.

A. Proof of Theorem 1

Two lemmas are needed to complete the proof of Theorem 1.

Lemma 1: Let $\{(\omega^{mJ}, V^{mJ})\}$ be the weight sequence generated by the online BP training procedure (1) with dynamic learning rate (2)–(4). Under Assumption 2, there holds

$$\begin{aligned}&\left\|(\omega^{(m+1)J}, V^{(m+1)J}) - (\omega, V)\right\|^2 \\ \le\ &\left\|(\omega^{mJ}, V^{mJ}) - (\omega, V)\right\|^2 \\ &-2\eta_m(E(\omega^{mJ}, V^{mJ}) - E(\omega, V)) + C^2\eta_m^2\end{aligned}$$

for any $(\omega, V) \in R^n \times R^{n\times p}$ and $m = 0, 1, ...$

$$\begin{aligned}&\left\| \left(\omega^{(m+1)J}, V^{(m+1)J}\right) - (\omega, V) \right\|^2 \\ \le\ &\left\| \left(\omega^{mJ}, V^{mJ}\right) - (\omega, V) \right\|^2 - 2\eta_m \sum_{j=0}^{J-1} \left(E_j\left(\omega^{mJ+j}, V^{mJ+j}\right) - E_j\left(\omega, V\right)\right) + \eta_m^2 \sum_{j=0}^{J-1} C_j^2 \\ =\ &\left\| \left(\omega^{mJ}, V^{mJ}\right) - (\omega, V) \right\|^2 - 2\eta_m \left(E\left(\omega^{mJ}, V^{mJ}\right) - E\left(\omega, V\right)\right) + \eta_m^2 \sum_{j=0}^{J-1} C_j^2 \\ &+2\eta_m \sum_{j=0}^{J-1} \left(E_j\left(\omega^{mJ}, V^{mJ}\right) - E_j\left(\omega^{mJ+j}, V^{mJ+j}\right)\right) \\ \le\ &\left\| \left(\omega^{mJ}, V^{mJ}\right) - (\omega, V) \right\|^2 - 2\eta_m \left(E\left(\omega^{mJ}, V^{mJ}\right) - E\left(\omega, V\right)\right) \\ &+2\eta_m \sum_{j=0}^{J-1} \left\langle \nabla E_j\left(\omega^{mJ}, V^{mJ}\right), \left(\omega^{mJ}, V^{mJ}\right) - \left(\omega^{mJ+j}, V^{mJ+j}\right)\right\rangle + \eta_m^2 \sum_{j=0}^{J-1} C_j^2 \\ \le\ &\left\| \left(\omega^{mJ}, V^{mJ}\right) - (\omega, V) \right\|^2 - 2\eta_m \left(E\left(\omega^{mJ}, V^{mJ}\right) - E\left(\omega, V\right)\right) \\ &+2\eta_m \sum_{j=0}^{J-1} \left\| \nabla E_j\left(\omega^{mJ}, V^{mJ}\right) \right\| \left\| \left(\omega^{mJ}, V^{mJ}\right) - \left(\omega^{mJ+j}, V^{mJ+j}\right) \right\| + \eta_m^2 \sum_{j=0}^{J-1} C_j^2 \\ \le\ &\left\| \left(\omega^{mJ}, V^{mJ}\right) - (\omega, V) \right\|^2 - 2\eta_m \left(E\left(\omega^{mJ}, V^{mJ}\right) - E\left(\omega, V\right)\right) \\ &+2\eta_m \sum_{j=1}^{J-1} C_j \left\| \sum_{i=0}^{j-1} \eta_m \nabla E_i\left(\omega^{mJ+i}, V^{mJ+i}\right) \right\| + \eta_m^2 \sum_{j=0}^{J-1} C_j^2 \\ \le\ &\left\| \left(\omega^{mJ}, V^{mJ}\right) - (\omega, V) \right\|^2 - 2\eta_m \left(E\left(\omega^{mJ}, V^{mJ}\right) - E\left(\omega, V\right)\right) + \eta_m^2 \left[2\sum_{j=1}^{J-1} C_j \left(\sum_{i=0}^{j-1} C_i\right) + \sum_{j=0}^{J-1} C_j^2\right] \\ =\ &\left\| \left(\omega^{mJ}, V^{mJ}\right) - (\omega, V) \right\|^2 - 2\eta_m \left(E\left(\omega^{mJ}, V^{mJ}\right) - E\left(\omega, V\right)\right) + \eta_m^2 \left(\sum_{j=1}^{J-1} C_j\right)^2 \\ =\ &\left\| \left(\omega^{mJ}, V^{mJ}\right) - (\omega, V) \right\|^2 - 2\eta_m \left(E\left(\omega^{mJ}, V^{mJ}\right) - E\left(\omega, V\right)\right) + \widetilde{C}^2\eta_m^2 \end{aligned} \tag{8}$$

Proof: According to Assumption 2 and the definition of the Frobenius norm, we have $\|\nabla E_j(\omega, V)\| = (\|\nabla E_{j,\omega}(\omega, V)\|^2 + \|\nabla E_{j,V}(\omega, V)\|^2)^{1/2} \le (C_{j,\omega}^2 + C_{j,V}^2)^{1/2}$. Write $C_j = (C_{j,\omega}^2 + C_{j,V}^2)^{1/2}$, then there hold

$$\|\nabla E_j(\omega, V)\| \le C_j, \quad j = 0, 1, \ldots, J-1. \tag{7}$$

In view of the online BP training formulation (1) and the boundedness inequality (7), we obtain that for any weight matrix $(\omega, V) \in R^n \times R^{n\times p}$

$$\begin{aligned}
&\left\| \left(\omega^{mJ+j+1}, V^{mJ+j+1}\right) - (\omega, V) \right\|^2 \\
&= \left\| \left(\omega^{mJ+j}, V^{mJ+j}\right) \right. \\
&\quad \left. -\eta_m \nabla E_j \left(\omega^{mJ+j}, V^{mJ+j}\right) - (\omega, V) \right\|^2 \\
&= \left\| \left(\omega^{mJ+j}, V^{mJ+j}\right) - (\omega, V) \right\|^2 \\
&\quad -2\eta_m \left\langle \nabla E_j \left(\omega^{mJ+j}, V^{mJ+j}\right), \left(\omega^{mJ+j}, V^{mJ+j}\right) - (\omega, V) \right\rangle \\
&\quad +\eta_m^2 \left\| \nabla E_j \left(\omega^{mJ+j}, V^{mJ+j}\right) \right\|^2 \\
&\le \left\| \left(\omega^{mJ+j}, V^{mJ+j}\right) - (\omega, V) \right\|^2 \\
&\quad -2\eta_m \left(E_j \left(\omega^{mJ+j}, V^{mJ+j}\right) - E_j(\omega, V)\right) + \eta_m^2 C_j^2
\end{aligned}$$

where the last inequality is followed by the convexity of $E_j(\omega, V)$. For any $(\omega, V) \in R^n \times R^{n\times p}$ and $m = 0, 1, \ldots$, by adding the above inequalities over $j = 0, 1, \ldots, J-1$, we can obtain (8) in Page 5, where $\widetilde{C} = \sum_{j=1}^{J-1} C_j$. Let $C \ge \widetilde{C}$ and then Lemma 1 follows.

Among other things, Lemma 1 guarantees that given some other point $(\omega, V) \in R^n \times R^{n\times p}$ with lower cost than (ω^{mJ}, V^{mJ}) and the current iterate (ω^{mJ}, V^{mJ}), the next iterate $(\omega^{(m+1)J}, V^{(m+1)J})$ will be closer to (ω, V) than (ω^{mJ}, V^{mJ}), provided the step-size η_m is sufficiently small (less than $(2(E(\omega^{mJ}, V^{mJ}) - E(\omega, V))/C^2)$) [28]. This fact is used repeatedly, with a variety of choices for (ω, V), in what follows.

Next lemma reveals a fundamental convergence result about the error sequence in the light of Lemma 1.

Lemma 2: Let $\{(\omega^{mJ}, V^{mJ})\}$ be the weight sequence generated by the online BP training procedure (1) with dynamic learning rate (2)–(4). Then under Assumptions 1 and 2, we have

$$\underline{\lim}_{m\to\infty} E(\omega^{mJ}, V^{mJ}) = E^*.$$

Proof: In order to arrive at a contradiction, assume that $\underline{\lim}_{m\to\infty} E(\omega^{mJ}, V^{mJ}) \ne E^*$, that is, $\underline{\lim}_{m\to\infty} E(\omega^{mJ}, V^{mJ}) > E^*$.

By virtue of $\delta_m \to 0^+$, we know that for any positive number ε with $0 < \varepsilon < \underline{\lim}_{m\to\infty} E(\omega^{mJ}, V^{mJ}) - E^*$, there is a sufficiently large number $\overline{m}$ and a point $(\overline{\omega}, \overline{V})$ such that for all $m \ge \overline{m}$

$$\delta_m < \varepsilon$$

and

$$\underline{\lim}_{m\to\infty} E\left(\omega^{mJ}, V^{mJ}\right) - E\left(\overline{\omega}, \overline{V}\right) \ge \varepsilon.$$

Hence

$$\begin{aligned}
E_m^{lev} &= \min_{\substack{0\le i\le m-1 \\ 0\le j\le J-1}} E\left(\omega^{iJ+j+1}, V^{iJ+j+1}\right) - \delta_m \\
&> \min_{\substack{0\le i\le m-1 \\ 0\le j\le J-1}} E\left(\omega^{iJ+j+1}, V^{iJ+j+1}\right) - \varepsilon \\
&\ge \underline{\lim}_{m\to\infty} E\left(\omega^{mJ}, V^{mJ}\right) - \varepsilon \\
&\ge E\left(\overline{\omega}, \overline{V}\right).
\end{aligned}$$

By using this relationship, the definition of η_m and Lemma 1 with $(\omega, V) = (\overline{\omega}, \overline{V})$, we obtain for any $m \ge \overline{m}$

$$\begin{aligned}
&\left\| \left(\omega^{(m+1)J}, V^{(m+1)J}\right) - \left(\overline{\omega}, \overline{V}\right) \right\|^2 \\
&\le \left\| \left(\omega^{mJ}, V^{mJ}\right) - \left(\overline{\omega}, \overline{V}\right) \right\|^2 \\
&\quad -2\eta_m \left(E\left(\omega^{mJ}, V^{mJ}\right) - E\left(\overline{\omega}, \overline{V}\right)\right) + C^2\eta_m^2 \\
&= \left\| \left(\omega^{mJ}, V^{mJ}\right) - \left(\overline{\omega}, \overline{V}\right) \right\|^2 \\
&\quad -2\beta_m \frac{E\left(\omega^{mJ}, V^{mJ}\right) - E_m^{lev}}{C^2} \\
&\quad \times \left[E\left(\omega^{mJ}, V^{mJ}\right) - E\left(\overline{\omega}, \overline{V}\right)\right] \\
&\quad +C^2 \left(\beta_m \frac{E\left(\omega^{mJ}, V^{mJ}\right) - E_m^{lev}}{C^2}\right)^2 \\
&\le \left\| \left(\omega^{mJ}, V^{mJ}\right) - \left(\overline{\omega}, \overline{V}\right) \right\|^2 \\
&\quad -2\beta_m \frac{\left(E\left(\omega^{mJ}, V^{mJ}\right) - E_m^{lev}\right)^2}{C^2} \\
&\quad +\beta_m^2 \frac{\left(E\left(\omega^{mJ}, V^{mJ}\right) - E_m^{lev}\right)^2}{C^2} \\
&= \left\| \left(\omega^{mJ}, V^{mJ}\right) - \left(\overline{\omega}, \overline{V}\right) \right\|^2 \\
&\quad -\beta_m (2-\beta_m) \frac{\left(E\left(\omega^{mJ}, V^{mJ}\right) - E_m^{lev}\right)^2}{C^2} \\
&\le \left\| \left(\omega^{mJ}, V^{mJ}\right) - \left(\overline{\omega}, \overline{V}\right) \right\|^2 - \underline{\beta}\left(2-\overline{\beta}\right) \frac{\delta_m^2}{C^2}.
\end{aligned} \tag{9}$$

By summing these inequalities over $m \ge \overline{m}$ and using the assumption that $\sum_{m=0}^{\infty} \delta_m^2 = \infty$, we have

$$\begin{aligned}
&\sum_{m=\overline{m}}^{\infty} \left(\left\| \left(\omega^{(m+1)J}, V^{(m+1)J}\right) - \left(\overline{\omega}, \overline{V}\right) \right\|^2 \right. \\
&\qquad \left. - \left\| \left(\omega^{mJ}, V^{mJ}\right) - \left(\overline{\omega}, \overline{V}\right) \right\|^2 \right) \\
&\le -\sum_{m=\overline{m}}^{\infty} \underline{\beta}\left(2-\overline{\beta}\right) \frac{\delta_m^2}{C^2} = -\infty.
\end{aligned} \tag{10}$$

On the other hand, since

$$\left\| \left(\omega^{(\overline{m}+k)J}, V^{(\overline{m}+k)J}\right) - \left(\overline{\omega}, \overline{V}\right) \right\|^2 \ge 0$$

for any positive integer k, we can deduce that

$$\begin{aligned}&\sum_{m=\overline{m}}^{\overline{m}+k-1}\left(\left\|\left(\omega^{(m+1)J},V^{(m+1)J}\right)-(\overline{\omega},\overline{V})\right\|^2\right.\\&\qquad\left.-\left\|\left(\omega^{mJ},V^{mJ}\right)-(\overline{\omega},\overline{V})\right\|^2\right)\\&=\left\|\left(\omega^{(\overline{m}+k)J},V^{(\overline{m}+k)J}\right)-(\overline{\omega},\overline{V})\right\|^2\\&\quad-\left\|\left(\omega^{\overline{m}J},V^{\overline{m}J}\right)-(\overline{\omega},\overline{V})\right\|^2\\&\geq-\left\|\left(\omega^{\overline{m}J},V^{\overline{m}J}\right)-(\overline{\omega},\overline{V})\right\|^2\end{aligned}$$

holds for any k, and then we have

$$\begin{aligned}&\sum_{m=\overline{m}}^{\infty}\left(\left\|\left(\omega^{(m+1)J},V^{(m+1)J}\right)-(\overline{\omega},\overline{V})\right\|^2\right.\\&\qquad\left.-\left\|\left(\omega^{mJ},V^{mJ}\right)-(\overline{\omega},\overline{V})\right\|^2\right)\\&\geq-\left\|\left(\omega^{\overline{m}J},V^{\overline{m}J}\right)-(\overline{\omega},\overline{V})\right\|^2.\end{aligned}\tag{11}$$

Combining (10) with (11), we immediately obtain

$$\left\|\left(\omega^{\overline{m}J},V^{\overline{m}J}\right)-(\overline{\omega},\overline{V})\right\|^2=\infty$$

which is obviously impossible, since for any given integer m, $\|(\omega^{mJ},V^{mJ})-(\overline{\omega},\overline{V})\|^2$ must be a finite value. This fact arrives at a contradiction, and consequently, we proved that $\underline{\lim}_{m\to\infty}E(\omega^{mJ},V^{mJ})=E^*$.

Proof of Theorem 1: Without loss of generality, we first assume that the minimum point of $E(\omega,V)$ is unique, i.e., there is only one minimum point $(\omega^*,V^*)\in\Omega^*$ such that $E(\omega^*,V^*)=E^*$. In this case, Theorem 1 does hold if we prove that given any $\varepsilon_0>0$, there exists a positive number M_0 such that $\|(\omega^{mJ},V^{mJ})-(\omega^*,V^*)\|<\varepsilon_0$ for all $m\geq M_0$.

In view of the continuity of the error function E and the fact that $\delta_m\to 0$, we can obtain that given $\varepsilon_1>0$ with $-(\underline{\beta}+\overline{\beta}^2)\varepsilon_1^2/C^2<(\varepsilon_0^2/4)$, there exist a sufficiently small number δ_0 with $0<\delta_0<(\varepsilon_0/2)$ and a sufficiently large number $M_1>0$ such that

$$E\left(\omega^{mJ},V^{mJ}\right)-E^*<\frac{\varepsilon_1}{2},\ \delta_m<\frac{\varepsilon_1}{2}$$

and hence

$$\begin{aligned}&E\left(\omega^{mJ},V^{mJ}\right)-E_m^{lev}\\&\leq E\left(\omega^{mJ},V^{mJ}\right)-E^*+\delta_m<\varepsilon_1\end{aligned}$$

for all $m\geq M_1$ with $\|(\omega^{mJ},V^{mJ})-(\omega^*,V^*)\|<\delta_0$.

We distinguish two cases in the following.

Case 1: $\|(\omega^{mJ},V^{mJ})-(\omega^*,V^*)\|<\delta_0$. In this case, if the next iterate satisfies

$$\begin{aligned}&\left\|\left(\omega^{(m+1)J},V^{(m+1)J}\right)-\left(\omega^*,V^*\right)\right\|\\&\leq\left\|\left(\omega^{mJ},V^{mJ}\right)-\left(\omega^*,V^*\right)\right\|\end{aligned}$$

then for such m we obviously have

$$\left\|\left(\omega^{(m+1)J},V^{(m+1)J}\right)-\left(\omega^*,V^*\right)\right\|<\delta_0.\tag{12}$$

Otherwise, by using Lemma 1 with $(\omega,V)=(\omega^*,V^*)$, we obtain that

$$\begin{aligned}&\left\|\left(\omega^{(m+1)J},V^{(m+1)J}\right)-\left(\omega^*,V^*\right)\right\|^2\\&-\left\|\left(\omega^{mJ},V^{mJ}\right)-\left(\omega^*,V^*\right)\right\|^2\\&\leq-2\beta_m\frac{E\left(\omega^{mJ},V^{mJ}\right)-E_m^{lev}}{C^2}\\&\qquad\times\left[E\left(\omega^{mJ},V^{mJ}\right)-E\left(\omega^*,V^*\right)\right]\\&\quad+C^2\left(\beta_m\frac{E\left(\omega^{mJ},V^{mJ}\right)-E_m^{lev}}{C^2}\right)^2\\&=-\frac{2\beta_m}{C^2}\left[E\left(\omega^{mJ},V^{mJ}\right)-E_m^{lev}\right]\left[E\left(\omega^{mJ},V^{mJ}\right)-E^*\right]\\&\quad+\frac{\beta_m^2}{C^2}\left[E\left(\omega^{mJ},V^{mJ}\right)-E_m^{lev}\right]^2\\&<-\frac{2\underline{\beta}}{C^2}\cdot\varepsilon_1\cdot\frac{\varepsilon_1}{2}+\frac{\overline{\beta}^2}{C^2}\cdot\varepsilon_1^2<\frac{\varepsilon_0^2}{4}.\end{aligned}\tag{13}$$

Then we have

$$\begin{aligned}&\left\|\left(\omega^{(m+1)J},V^{(m+1)J}\right)-\left(\omega^*,V^*\right)\right\|^2\\&<\left\|\left(\omega^{mJ},V^{mJ}\right)-\left(\omega^*,V^*\right)\right\|^2+\frac{\varepsilon_0^2}{4}\\&\leq\left(\left\|\left(\omega^{mJ},V^{mJ}\right)-\left(\omega^*,V^*\right)\right\|+\frac{\varepsilon_0}{2}\right)^2\end{aligned}$$

and hence

$$\begin{aligned}&\left\|\left(\omega^{(m+1)J},V^{(m+1)J}\right)-\left(\omega^*,V^*\right)\right\|\\&\leq\left\|\left(\omega^{mJ},V^{mJ}\right)-\left(\omega^*,V^*\right)\right\|+\frac{\varepsilon_0}{2}\\&<\delta_0+\frac{\varepsilon_0}{2}<\varepsilon_0.\end{aligned}\tag{14}$$

Combining with (12) and (14), we then conclude the first property as follows:

$$\begin{aligned}&(*)\ \text{when}\left\|\left(\omega^{mJ},V^{mJ}\right)-\left(\omega^*,V^*\right)\right\|<\delta_0\\&\quad\text{and } m\geq M_1,\text{there must hold}\\&\left\|\left(\omega^{(m+1)J},V^{(m+1)J}\right)-\left(\omega^*,V^*\right)\right\|<\varepsilon_0.\end{aligned}$$

Case 2: $\delta_0\leq\|(\omega^{mJ},V^{mJ})-(\omega^*,V^*)\|\leq\varepsilon_0$. Denote by $E_0=\{E(\omega,V):\delta_0\leq\|(\omega^{mJ},V^{mJ})-(\omega^*,V^*)\|\leq\varepsilon_0\}$, then due to the uniqueness of (ω^*,V^*) and Assumption 1, we know that there is no minimum point in E_0. And further, by the compactness of E_0, we have $\inf E_0-E^*=a>0$. Therefore, through comparing the two items $(2\beta_m/C^2)[E(\omega^{mJ},V^{mJ})-E_m^{lev}][E(\omega^{mJ},V^{mJ})-E^*]$ and $(\beta_m^2/C^2)\cdot[E(\omega^{mJ},V^{mJ})-E_m^{lev})]^2$ in (13), there is a sufficiently large number $M_2>0$ such that

$$\begin{aligned}&\frac{\frac{2\beta_m}{C^2}\left[E\left(\omega^{mJ},V^{mJ}\right)-E_m^{lev}\right]\left[E\left(\omega^{mJ},V^{mJ}\right)-E^*\right]}{\frac{\beta_m^2}{C^2}\left[E\left(\omega^{mJ},V^{mJ}\right)-E_m^{lev}\right]^2}\\&\geq\frac{2}{\overline{\beta}}\frac{E\left(\omega^{mJ},V^{mJ}\right)-E^*}{E\left(\omega^{mJ},V^{mJ}\right)-E^*+\delta_m}>\frac{2}{\overline{\beta}}\frac{a}{a+\delta_m}>1\end{aligned}$$

for all $m \geq M_2$ because $0 < \beta_m \leq \overline{\beta} < 2$ and $\delta_m \to 0$. Hence by inequality (13), we deduce that

$$\left\| \left(\omega^{(m+1)J}, V^{(m+1)J}\right) - \left(\omega^*, V^*\right) \right\| \leq \left\| \left(\omega^{mJ}, V^{mJ}\right) - \left(\omega^*, V^*\right) \right\|$$

and then we obtain the second property

$$(**) \text{ when } \delta_0 \leq \left\| \left(\omega^{mJ}, V^{mJ}\right) - \left(\omega^*, V^*\right) \right\| < \varepsilon_0$$
and $m \geq M_2$, there must hold
$$\left\| \left(\omega^{(m+1)J}, V^{(m+1)J}\right) - \left(\omega^*, V^*\right) \right\| \leq \left\| \left(\omega^{mJ}, V^{mJ}\right) - \left(\omega^*, V^*\right) \right\|.$$

According to Assumption 1, the uniqueness of (ω^*, V^*) and the fact that $\underline{\lim}_{m\to\infty} E(\omega^{mJ}, V^{mJ}) = E^*$ from Lemma 2, we know that there exists a sufficiently large number M_0 with $M_0 \geq \max\{M_1, M_2\}$ such that

$$\left\| \left(\omega^{M_0 J}, V^{M_0 J}\right) - \left(\omega^*, V^*\right) \right\| < \delta_0.$$

Now let $m = M_0 + k$.

1) If $k = 0$, then there obviously holds

$$\left\| \left(\omega^{M_0 J}, V^{M_0 J}\right) - \left(\omega^*, V^*\right) \right\| < \delta_0 < \varepsilon_0.$$

2) Suppose that

$$\left\| \left(\omega^{(M_0+k)J}, V^{(M_0+k)J}\right) - \left(\omega^*, V^*\right) \right\| < \varepsilon_0.$$

Combining the property $(*)$, which holds when $\|(\omega^{(M_0+k)J}, V^{(M_0+k)J}) - (\omega^*, V^*)\| < \delta_0$ (Case 1 above), with the property $(**)$, which holds when $\delta_0 \leq \|(\omega^{(M_0+k)J}, V^{(M_0+k)J}) - (\omega^*, V^*)\| < \varepsilon_0$ (Case 2 above), we see that there always holds

$$\left\| \left(\omega^{(M_0+k+1)J}, V^{(M_0+k+1)J}\right) - \left(\omega^*, V^*\right) \right\| < \varepsilon_0.$$

Therefore, from 1) and 2), we show that

$$\left\| \left(\omega^{mJ}, V^{mJ}\right) - \left(\omega^*, V^*\right) \right\| < \varepsilon_0$$

for all $m \geq M_0$ by induction on k. This implies that $\lim_{m\to\infty}(\omega^{mJ}, V^{mJ}) = (\omega^*, V^*)$.

Next, let us assume that $E(\omega, V)$ has finite number of minimum points, denoted by (ω_1^*, V_1^*), $(\omega_2^*, V_2^*), \ldots, (\omega_k^*, V_k^*)$, respectively. According to $\underline{\lim}_{m\to\infty} E(\omega^{mJ}, V^{mJ}) = E^*$, we know that there must exist a minimum point (ω_i^*, V_i^*) $(i \in \{1, 2, \ldots, k\})$ such that (ω_i^*, V_i^*) is a limit point of the weight sequence $\{(\omega^{mJ}, V^{mJ})\}$. Then with (ω_i^*, V_i^*), using the same proof shown above under the unique-minimum-point assumption, we can prove that

$$\lim_{m\to\infty} \left(\omega^{mJ}, V^{mJ}\right) = \left(\omega_i^*, V_i^*\right).$$

The proof of Theorem 1 is thus completed.

B. Proof of Theorem 2

Assumption 1 obviously implies that Ω^* must be bounded. Since Ω^* is also closed due to the continuity of E, we can then conclude that

Lemma 3: The set Ω^* is compact under Assumptions 1 and 2.

The following lemma lists several fundamental properties of compact sets which are useful for our further discussion.

Lemma 4: Let $X, Y \subset R^n$ be two compact sets.

1) If X and Y are disjoint, then there exist two points $x_0 \in X$, $y_0 \in Y$ such that $\rho(x_0, y_0) = \rho(X, Y) > 0$ where $\rho(X, Y) = \inf_{\substack{x\in X \\ y\in Y}} \rho(x, y)$. In the particular case where Y is a single-point set, denoted by $Y = \{y\}$, there exists a point $x_0 \in X$ such that $\rho(y, x_0) = \rho(y, X)$.

2) For any open set $G \supset X$, there exists an $(1/n)$−neighborhood $U_{(1/n)}(X) = \{x : \exists z \in X, \rho(x, z) < (1/n)\}$ of X such that $U_{(1/n)}(X) \subset G$.

Proof: 1) Denote by $\rho(X, Y) = \inf_{\substack{x\in X \\ y\in Y}} \rho(x, y) = a$. Then for any positive integer n, there exist two sequences $\{x_n\} \subset X$ and $\{y_n\} \subset Y$ such that $a \leq \rho(x_n, y_n) < a + (1/n)$. Since X and Y are two disjoint compact sets, we know that there have subsequences $\{x_{n_k}\} \subset \{x_n\}$ and $\{y_{n_k}\} \subset \{y_n\}$ such that $x_{n_k} \to x_0$ and $y_{n_k} \to y_0$ where $x_0 \in X$, $y_0 \in Y$ with $x_0 \neq y_0$. Hence there holds $a \leq \rho(x_{n_k}, y_{n_k}) < a + (1/n)$. By the continuity of ρ, we then derive that $\rho(x_0, y_0) = a$. Therefore, we obtain that $\rho(X, Y) = \rho(x_0, y_0) > 0$.

It is obvious that Y is compact in the case where $Y = \{y\}$. Hence, as a corollary of the result we proved above, there exists a point $x_0 \in X$ such that $\rho(y, x_0) = \rho(y, X)$.

2) Since $X \subset G$, for any point $x \in X$, there exists a positive number r_x such that $S(x, r_x) \subset G$ where $S(x, r_x)$ is a sphere with center x and radius r_x. Thus we know that $\Gamma = \{S(x, (1/2)r_x) : x \in X\}$ is an open covering of X. Since X is also compact, we know that there have finite elements $S(x_1, (1/2)r_{x_1}), S(x_2, (1/2)r_{x_2}), \ldots, S(x_k, (1/2)r_{x_k})$ in Γ covering X.

Denote $r^* = \min\{(1/2)r_{x_1}, (1/2)r_{x_2}, \ldots, (1/2)r_{x_k}\}$, then for any $x \in U_{r^*}(X)$, there exists a point $z \in X$ such that $\rho(x, z) < r^*$. Meanwhile, there must exist a sphere $S(x_l, (1/2)r_{x_l})$ containing the point z where $l \in \{1, 2, \ldots k\}$. Therefore, we have $\rho(x, x_l) \leq \rho(x, z) + \rho(z, x_l) < r^* + (1/2)r_{x_l} < r_{x_l}$ which shows that $x \in S(x_l, r_{x_l}) \subset G$. This obviously reveals that $U_{r^*}(X) \subset G$. Let n be a positive integer with $(1/n) < r^*$ and then 2) of Lemma 4 follows.

In view of Lemmas 3 and 4, and Theorem 1, we can prove Theorem 2 as follows.

Proof of Theorem 2: Since Ω^* is compact, by Lemma 4 1), for any point (ω^{mJ}, V^{mJ}) in the sequence $\{(\omega^{mJ}, V^{mJ})\}$ generated by the online BP training procedure with the dynamic learning rate, there exists a point $(\omega_0^{mJ}, V_0^{mJ}) \in \Omega^*$ such that

$$\rho\left(\left(\omega^{mJ}, V^{mJ}\right), \Omega^*\right) = \rho\left(\left(\omega^{mJ}, V^{mJ}\right), \left(\omega_0^{mJ}, V_0^{mJ}\right)\right).$$

Using Lemma 1 with $(\omega_0^{mJ}, V_0^{mJ})$ yields

$$\rho^2\left(\left(\omega^{(m+1)J}, V^{(m+1)J}\right), \left(\omega_0^{mJ}, V_0^{mJ}\right)\right) = \left\| \left(\omega^{(m+1)J}, V^{(m+1)J}\right) - \left(\omega_0^{mJ}, V_0^{mJ}\right) \right\|^2$$

$$\begin{aligned}
&\le \left\| \left(\omega^{mJ}, V^{mJ}\right) - \left(\omega_0^{mJ}, V_0^{mJ}\right) \right\|^2 \\
&\quad -2\eta_m \left(E\left(\omega^{mJ}, V^{mJ}\right) - E\left(\omega_0^{mJ}, V_0^{mJ}\right)\right) + C^2\eta_m^2 \\
&= \rho^2\left(\left(\omega^{mJ}, V^{mJ}\right), \Omega^*\right) \\
&\quad -2\eta_m \left(E\left(\omega^{mJ}, V^{mJ}\right) - E\left(\omega_0^{mJ}, V_0^{mJ}\right)\right) + C^2\eta_m^2.
\end{aligned}$$

Then according to the definition of $\rho((\omega^{mJ}, V^{mJ}), \Omega^*)$ and the fact that $E(\omega_0^{mJ}, V_0^{mJ}) = E^*$, we can immediately deduce that

$$\begin{aligned}
&\rho^2\left(\left(\omega^{(m+1)J}, V^{(m+1)J}\right), \Omega^*\right) \\
&\le \rho^2\left(\left(\omega^{(m+1)J}, V^{(m+1)J}\right), \left(\omega_0^{mJ}, V_0^{mJ}\right)\right) \\
&\le \rho^2\left(\left(\omega^{mJ}, V^{mJ}\right), \Omega^*\right) \\
&\quad -2\eta_m \left(E\left(\omega^{mJ}, V^{mJ}\right) - E^*\right) + C^2\eta_m^2 \\
&= \rho^2\left(\left(\omega^{mJ}, V^{mJ}\right), \Omega^*\right) \\
&\quad -\frac{2\beta_m}{C^2}\left[E\left(\omega^{mJ}, V^{mJ}\right) - E_m^{lev}\right]\left[E\left(\omega^{mJ}, V^{mJ}\right) - E^*\right] \\
&\quad + \frac{\beta_m^2}{C^2}\left[E\left(\omega^{mJ}, V^{mJ}\right) - E_m^{lev}\right]^2. \qquad (15)
\end{aligned}$$

In the light of (15), similar to the proof of Theorem 1, we then conclude that given any $\varepsilon > 0$, there exists a number δ with $0 < \delta < \varepsilon$ such that the following two properties hold: $(*)$ $\rho((\omega^{(m+1)J}, V^{(m+1)J}), \Omega^*) < \varepsilon$ for sufficiently large m as $\rho((\omega^{mJ}, V^{mJ}), \Omega^*) < \delta$, $(**)$ $\rho((\omega^{(m+1)J}, V^{(m+1)J}), \Omega^*) \le \rho((\omega^{mJ}, V^{mJ}), \Omega^*)$ for sufficiently large m as $\delta \le \rho((\omega^{mJ}, V^{mJ}), \Omega^*) < \varepsilon$.

Moreover, according to Assumption 1, the compactness of Ω^* and the fact that $\underline{\lim}_{m\to\infty} E(\omega^{mJ}, V^{mJ}) = E^*$ from Lemma 2, we obtain that there exists a sufficiently large number $M_0 > 0$ such that $\rho((\omega^{M_0 J}, V^{M_0 J}), \Omega^*) < \delta$. Then by virtue of the above two properties $(*)$ and $(**)$, we can draw a conclusion that $\rho((\omega^{mJ}, V^{mJ}), \Omega^*) < \varepsilon$ always hold for all $m \ge M_0$ by induction on m. This implies that

$$\lim_{m\to\infty} \rho\left(\left(\omega^{mJ}, V^{mJ}\right), \Omega^*\right) = 0.$$

C. Proof of Theorem 3

With the above results, Theorem 3 can be shown directly as follows.

Proof of Theorem 3: According to Theorem 2 and the continuity of E, given $\varepsilon > 0$, for any point (ω^*, V^*) in Ω^*, there exists a neighborhood U^* of (ω^*, V^*) such that $E(\omega, V) < E^* + \varepsilon$ for all $(\omega, V) \in U^*$. Then $G = \bigcup U^*$ is an open set containing Ω^* as a subset and $E(\omega, V) < E^* + \varepsilon$ holds for all $(\omega, V) \in G$. By using Lemma 4 2), we can then conclude that there is a neighborhood $U_{1/m}(\Omega^*)$ such that $U_{1/m}(\Omega^*) \subset G$. That is to say, there is a positive number M such that $E(\omega^{mJ}, V^{mJ}) < E^* + \varepsilon$ for all $m \ge M$. The proof of Theorem 3 is then completed.

D. Proof of Theorem 4

We finally prove Theorem 4 with other two useful lemmas.

Lemma 5: Suppose that the set X is nonempty, closed and connected. Then X has either single point (i.e., $X = \{x\}$) or continuum cardinal number (i.e., $\bar{X} = \aleph$).

Proof: We only need to prove that X must have continuum cardinal number if X is not a single-point set.

Under this assumption, we first prove that X has no isolated point. In order to arrive at a contradiction, assume that x_0 is an isolated point of X, then $X - \{x_0\}$ must be nonempty. Let $A_1 = \{x_0\}$ and $A_2 = X - \{x_0\}$, then $X = A_1 \cup A_2$, $\bar{A}_1 \cap A_2 = \emptyset$ and $\bar{A}_2 \cap A_1 = \emptyset$ where $\bar{A}_i$ denotes the closure of A_i $(i = 1, 2)$. This shows that X is not connected which is obviously contradict to the assumption on the connectivity of X. Therefore, X must have no isolated point.

Then according to the fact that a nonempty and closed set must have continuum cardinal number if it has no isolated point, it is directly proved that X has continuum cardinal number, that is, $\bar{X} = \aleph$, justifying Lemma 5.

Lemma 6: The set W of all the limit points of the weight sequence $\{(\omega^{mJ}, V^{mJ})\}$ is closed, compact and connected under Assumptions 1 and 2.

Proof: It is obvious that $W \subset \Omega^*$ is closed and compact according to the definition of W and Theorem 3.

The connectivity of W can be proved by the contradictory method. Suppose that W is not connected, then W can be expressed as $W = W_1 \cup W_2$ where W_1, W_2 are two closed and compact subsets of W with $W_1 \ne \emptyset$, $W_2 \ne \emptyset$ and $W_1 \cap W_2 = \emptyset$. Consequently, there must have $\rho(W_1, W_2) > 0$ by Lemma 4 1).

Suppose that $\rho(W_1, W_2) = 3\varepsilon > 0$. Denote by U, U_1, U_2 the $\varepsilon-$neighborhoods of W, W_1, W_2, respectively, then $U = U_1 \cup U_2$, $\|(\omega_1, V_1) - (\omega_2, V_2)\| > \varepsilon$ for any $(\omega_1, V_1) \in U_1$ and $(\omega_2, V_2) \in U_2$ and there is a positive number M_1 such that $(\omega^{mJ}, V^{mJ}) \notin U^C$ for all $m \ge M_1$ where U^C denotes the complement of U.

Moreover, by virtue of (1) and $\eta_m \to 0$, there exists a positive number M_2 such that $\|(\omega^{(m+1)J}, V^{(m+1)J}) - (\omega^{mJ}, V^{mJ})\| \to 0$ for all $m \ge M_2$.

Let $M^* = \max\{M_1, M_2\}$, then there must exist a number m_0 with $m_0 \ge M^*$ such that $(\omega^{m_0 J}, V^{m_0 J}) \in U_1$ as well as $(\omega^{(m_0+1)J}, V^{(m_0+1)J}) \notin U_2$. Therefore, we can prove that $U_2 = \emptyset$ by induction on m, which is obviously a contradiction. This arrives Lemma 6.

With the above Lemmas 5 and 6, we can finally prove the following main result on the global strong convergence of the online BP training procedure.

Proof of Theorem 4: In the light of Lemma 5, we know that W is closed, compact and connected under Assumptions 1 and 2. Then by Lemma 6, we know that W has either single point or continuum cardinal number. However, Assumption 3 reveals that W cannot have continuum cardinal number, and hence, W can only be a single-point set. Then Theorem 4 follows directly from Theorem 1.

TABLE I
THREE-BIT PARITY PROBLEM

input			output	input			output
1	1	1	1	0	1	1	0
0	0	1	1	1	0	1	0
1	1	0	0	0	1	0	1
1	0	0	1	0	0	0	0

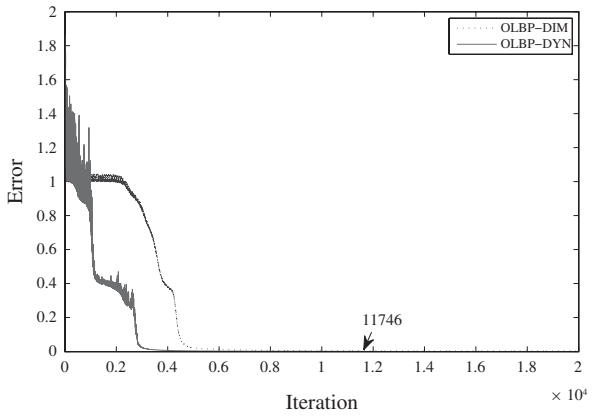


Fig. 1. Error curves of OLBP-DYN and OLBP-DIM for three-bit parity problem.

V. EXPERIMENTAL VERIFICATION

To illustrate the theoretical results we proved above, in this section, we investigate the performance of the online BP training procedure with the proposed dynamic learning rate (OLBP-DYN) and the online BP training procedure with the diminishing learning rate (OLBP-DIM, [25]) on two problems: 1) Three-bit parity problem; and 2) Sonar benchmark problem. All the simulations were carried out in MATLAB R2010b environment running on an ordinary PC with 3.2 GHz CPU. The sigmoidal activation function $g(x) = 1/(1 + \exp(-2x))$ is used for both hidden and output nodes of the following network architectures. In the proposed dynamic learning rate (2)–(4), β_m is randomly chosen in $(0, 2)$ based on uniform distribution and the parameter δ_m is set as $\delta_m = 1/\sqrt{m}$. For the case of OLBP-DIM, we set the diminishing learning rate as $\eta_m = 1/m$ which satisfies the conditions $\eta_m > 0$, $\sum_{m=1}^{\infty} \eta_m = \infty$ and $\sum_{m=1}^{\infty} \eta_m^2 < \infty$ (which are supposed in many literatures, such as in [25]).

A. Parity Problem

In this subsection, we consider the three-bit parity problem [32] to verify the convergence property of the online BP training with two different learning rates: OLBP-DYN and OLBP-DIM. The inputs and the target outputs of the training examples are shown in Table I.

The architecture of the neural network used in this example is 3-6-1 (input-hidden-output nodes). For both cases, the initial weights are randomly chosen in $[-0.5, 0.5]$ based on uniform distribution. For OLBP-DYN, $C = 1$. The training process will be terminated once the maximum number of iterations 20 000 is reached or the target error 0.001 is satisfied. In this simulation, the average results over 20 trials are obtained for both cases.

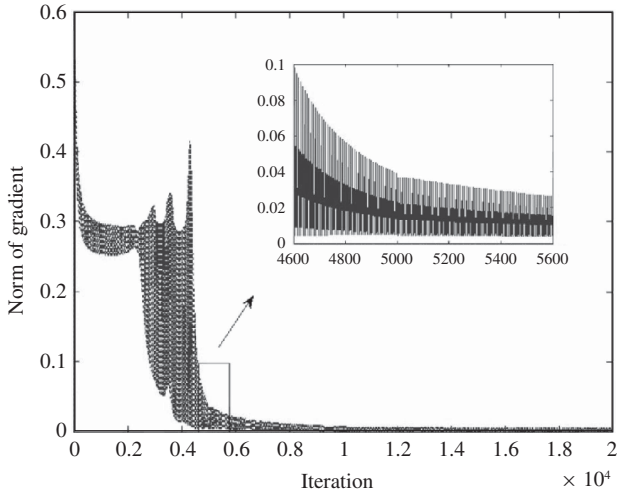


Fig. 2. Norm of gradient curve of OLBP-DIM for three-bit parity problem.

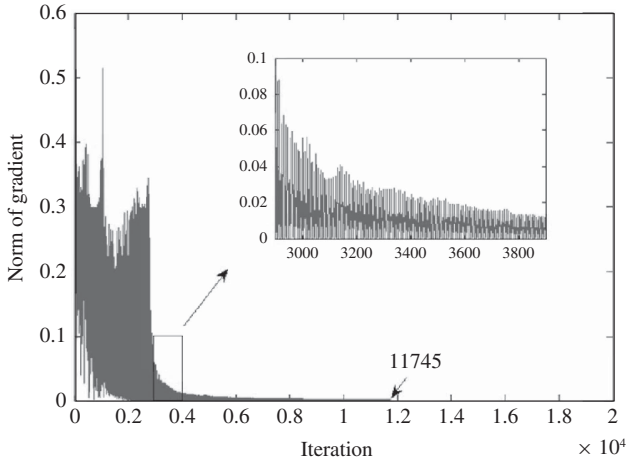


Fig. 3. Norm of gradient curve of OLBP-DYN for three-bit parity problem.

Fig. 1 shows the average error function decreasing trends of OLBP-DYN and OLBP-DIM. From Fig. 1, we observe that with the dynamic learning rate, the error decreases to zero at the 11746th epoch, whereas with the diminishing learning rate, the error decreases to 0.0033 after 20 000 epochs. Figs. 2 and 3 show the update of the norm of gradient of the error functions with the increase of the iteration for OLBP-DYN and OLBP-DIM, respectively. As observed from Figs. 2 and 3, the norm of gradient decreases below 0.1 around 2900th epoch in OLBP-DYN whereas around 4600th epoch in OLBP-DIM, which shows that the convergence speed of OLBP-DYN is over one and a half times faster than that of OLBP-DIM. Moreover, the norm of gradient attains zero at the 11745th epoch in OLBP-DYN and it only attains 0.0031 after 20 000 epochs in OLBP-DIM. These results obviously support the convergence analysis of this paper.

TABLE II
PERFORMANCE COMPARISON OF OLBP-DYN AND OLBP-DIM FOR SONAR PROBLEM

Algorithms	Training Time(s)	Error	Training Accuracy	Testing Accuracy
OLBP-DYN	34.84	**0.92**	**96.66%**	**86.88%**
OLBP-DIM [25]	**30.73**	1, 66	91.28%	80.26%

B. Sonar Problem

The Sonar benchmark problem is a well-known classification problem proposed in [33]. The date set is publicly available from UCI database [34], which comprises 208 input vectors, each with 60 components. In this simulation, 105 samples (56 from Metals and 49 from Rocks) are randomly chosen as the training data set and 101 samples (54 from Metals and 47 from Rocks) are randomly chosen as the testing data set. The network used in this example is 60-10-1. The initial values of the weights are randomly chosen in $[-2, 2]$ based on uniform distribution. In this example, $C = 5$ and the stop criteria are the maximum number of iterations 30 000, or the error reduced to 0.8. In this simulation, the average results over 10 trials are obtained.

Table II summarizes the average results for the Sonar problem in terms of the training time, error, training accuracy, and testing accuracy. As observed from Table II, after 30 000 iterations, OLBP-DYN reached the error 0.92 which is similar to the target error 0.8, whereas OLBP-DIM only attains the error 1.66. Meanwhile, OLBP-DYN exhibits a higher success rate than OLBP-DIM in both training process and testing process. These results also reveal that the proposed dynamic learning rate can obtain better performance than the diminishing learning rate, which further show the efficiency of the proposed dynamic learning rate.

VI. CONCLUSION

This paper proposed a new dynamic learning scheme implemented during the online BP training procedure for single-hidden layer feedforward networks. Our proposed dynamic learning rate is more reasonable and reliable in applications as it utilizes the error information during the training procedure. We then analyzed the global convergence of the online BP training procedure with the proposed dynamic learning rate in two aspects, with one claiming that the error sequence $\{E(\omega^{mJ+j}, V^{mJ+j})\}$ converges to the global minimum error (the globally weak convergence), and another concluding that the weight sequence $\{(\omega^{mJ+j}, V^{mJ+j})\}$ converges to a fixed point at which the error function attains its global minimum (the globally strong convergence) under weaker conditions. The obtained convergence results generalize those existing analyses (particularly, the results obtained recently by Xu *et al.* [25]) in the sense that: 1) they resolve the global convergence issue on the online BP training procedure, which was not involved in [25]; and 2) they are effective for a very general family of learning rates. Comparing with those used in [25], this paper provides a broader and more reasonable selection of the learning rates.

REFERENCES

[1] C. M. Bishop, *Neural Networks for Pattern Recognition*. London, U.K.: Oxford Univ. Press, 1995.

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

[3] S. Haykin, *Neural Networks: A Comprehensive Foundation*. New York: Macmillan, 1994.

[4] C. G. Looney, *Pattern Recognition Using Neural Networks: Theory and Algorithms for Engineers and Scientists*. New York: Oxford Univ. Press, 1997.

[5] D. Wang and P. Bao, "Enhancing the estimation of plant Jacobian for adaptive neural inverse control," *Neurocomputing*, vol. 34, nos. 1–4, pp. 99–115, Sep. 2000.

[6] D. Wang and X. Ma, "A hybrid image retrieval system with user's relevance feedback using neurocomputing," *Informatica*, vol. 29, no. 3, pp. 271–280, 2005.

[7] K. I.-J. Ho, C.-S. Leung, and J. Sum, "Convergence and objective functions of some fault/noise-injection-based online learning algorithms for RBF networks," *IEEE Trans. Neural Netw.*, vol. 21, no. 6, pp. 938–947, Jun. 2010.

[8] B. Chen, Y. Zhu, and J. Hu, "Mean-square converge analysis of ADALINE training with minimum error entropy criterion," *IEEE Trans. Neural Netw.*, vol. 21, no. 7, pp. 1168–1178, Jul. 2010.

[9] X. Wang and Y. Huang, "Converge study in extended Kalman filter-based training of recurrent neural networks," *IEEE Trans. Neural Netw.*, vol. 22, no. 4, pp. 588–600, Apr. 2011.

[10] L. T. Fine and S. Mukherjee, "Parameter convergence and learning curves for neural networks," *Neural Comput.*, vol. 11, no. 3, pp. 747–769, 1998.

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

[12] A. A. Gaivoronski, "Convergence properties of backpropagation for neural nets via theory of stochastic gradient methods. Part I," *Opt. Methods Softw.*, vol. 4, no. 2, pp. 117–134, 1994.

[13] C.-M. Kuan and K. Hornik, "Convergence of learning algorithms with constant learning rates," *IEEE Trans. Neural Netw.*, vol. 2, no. 5, pp. 484–489, Sep. 1991.

[14] H. J. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*. New York: Springer-Verlag, 2003.

[15] S.-H. Oh, "Improving the error BP algorithm with a modified error function," *IEEE Trans. Neural Netw.*, vol. 8, no. 3, pp. 799–803, May 1997.

[16] H. White, "Some asymptotic results for learning in single hidden-layer feedforward neural network models," *J. Amer. Stat. Assoc.*, vol. 84, no. 408, pp. 1003–1013, Dec. 1989.

[17] Z.-X. Li, W. Wu, G.-R. Feng, and H.-F. Lu, *Convergence of an Online Gradient Method for BP Neural Networks with Stochastic Inputs* (Lecture Notes in Computer Science). New York: Springer-Verlag, 2005, pp. 720–729.

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

[19] Z.-Q. Luo and P. Tseng, "Analysis of an approximate gradient projection method with application to the backpropagation algorithm," *Optim. Methods Softw.*, vol. 4, no. 2, pp. 85–101, 1994.

[20] O. L. Manggasarian and M. V. Solodov, "Serial and parallel backpropagation convergence via nonmonotone perturbed minimization," *Optim. Methods Softw.*, vol. 4, no. 2, pp. 103–116, 1994.

[21] W. Wu, G.-R. Feng, and X. Li, "Training multilayer perceptrons via minimization of sum of ridge functions," *Adv. Comput. Math.*, vol. 17, no. 4, pp. 331–347, 2002.

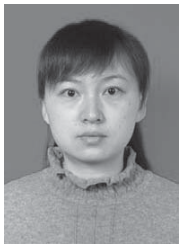
[22] W. Wu, G.-R. Feng, Z.-X. Li, and Y.-S. Xu, "Deterministic convergence of an online gradient method for BP neural networks," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 533–540, May 2005.

[23] W. Wu and Z.-Q. Shao, "Convergence of an online gradient methods for continuous perceptrons with linearly separable training patterns," *Appl. Math. Lett.*, vol. 16, no. 7, pp. 999–1002, 2003.

[24] W. Wu and Y.-S. Xu, "Deterministic convergence of an online gradient method for neural networks," *J. Comput. Appl. Math.*, vol. 144, no. 1, pp. 335–347, 2002.

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

[26] H. Zhang, W. Wu, F. Liu, and M. Yao, "Boundedness and converge of online gradient method with penalty for feedforward neural network," *IEEE Trans. Neural Netw.*, vol. 20, no. 6, pp. 1050–1054, Jun. 2009.
[27] 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.
[28] A. Nedic and D. P. Bertsekas, "Incremental subgradient methods for nondifferentiable optimization," *SIAM J. Optim.*, vol. 12, no. 1, pp. 109–138, 2001.
[29] S. Duffner and C. Garcia, "An online backpropagation algorithm with validation error-based adaptive learning rate," in *Proc. Int. Conf. Artif. Neural Netw.*, vol. 1. 2007, pp. 249–258.
[30] P. D. Bertsekas and J. N. Tsitsiklis, "Gradient convergence in gradient methods with errors," *SIAM J. Optim.*, vol. 10, no. 3, pp. 627–642, 2000.
[31] K. Kuratowski, *Introduction to Set Theory and Topology*. Warszawa, Poland: PWN Publishers, 1972.
[32] D. Liu, M. E. Hohil, and S. H. Smith, "N-bit parity neural networks: New solutions based on linear programming," *Neurocomputing*, vol. 48, nos. 1–4, pp. 477–488, Oct. 2002.
[33] R. P. Gorman and T. J. Sejnowski, "Analysis of hidden units in a layered network trained to classify sonar targets," vol. 1, no. 1, pp. 75–89, 1988.
[34] C. L. Black and C. J. Merz. (1998). *UCI Repository of Machine Learning Databases*. Dep. Inf. Comput. Sci., Univ. California, Irvine [Online]. Available: http://www.ics.uci.edu/∼mlearn/mlrepository.html



Rui Zhang received the B.Sc. and M.Sc. degrees in mathematics from Northwest University, Xi'an, China, in 1994 and 1997, respectively. She is currently a Ph.D. candidate at the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

She has been with the Department of Mathematics, Northwest University, since 1997, where she is now an Associate Professor. From August 2004 to January 2005, she was a Visiting Scholar with the Department of Mathematics, University of Illinois at Champaign-Urbana, Urbana. Her current research interests include extreme learning machines, machine learning, and neural networks.

Zongben Xu received the Ph.D. degree in mathematics from Xi'an Jiaotong University, Xi'an, China, in 1987.

He now serves as a Vice President of Xi'an Jiaotong University, the Chief Scientist of the National Basic Research Program of China (973 Project), and the Director of the Institute for Information and System Sciences of the university. His current research interests include nonlinear functional analysis and intelligent information processing.

Dr. Xu holds the National Natural Science Award of China in 2007 and is a winner of the CSIAM Su Buchin Applied Mathematics Prize in 2008. He delivered a 45-minute talk at the International Congress of Mathematicians 2010.

Guang-Bin Huang (M'98–SM'04) received the B.Sc. degree in applied mathematics and the M.Eng. degree in computer engineering from Northeastern University, Shenyang, China, in 1991 and 1994, respectively, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 1999. During undergraduate period, he also concurrently studied in the Applied Mathematics Department and the Wireless Communication Department, Northeastern University, China.

He was a Research Fellow with the Singapore Institute of Manufacturing Technology (formerly known as the Gintic Institute of Manufacturing Technology), Singapore, where he has led/implemented several key industrial projects from June 1998 to May 2001. In May 2001, he has been an Assistant Professor and Associate Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University. His current research interests include machine learning, computational intelligence, and extreme learning machines.

Dr. Huang serves as an Associate Editor of *Neurocomputing* and the IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS—PART B.

Dianhui Wang (SM'03) received the Ph.D. degree from Northeastern University, Shenyang, China, in 1995.

He is currently an Associate Professor with the Department of Computer Science and Computer Engineering, La Trobe University, Melbourne, Victoria, Australia. From 1995 to 2001, he was a Post-Doctoral Fellow with Nanyang Technological University, Singapore, and a Researcher with Hong Kong Polytechnic University, Hong Kong. He is associated with the State Key Laboratory of Synthetical Automation of Process Industries, Northeastern University. His current research interests include data mining and computational intelligence systems for bioinformatics, image processing, and engineering applications.