When Does Online BP Training Converge?

Zong-Ben Xu, Rui Zhang, and Wen-Feng Jing

***Abstract*—The backpropogation (BP) neural networks have been widely applied in scientific research and engineering. The success of the application, however, relies upon the convergence of the training procedure involved in the neural network learning. We settle down the convergence analysis issue through proving two fundamental theorems on the convergence of the online BP training procedure. One theorem claims that under mild conditions, the gradient sequence of the error function will converge to zero (the weak convergence), and another theorem concludes the convergence of the weight sequence defined by the procedure to a fixed value at which the error function attains its minimum (the strong convergence). The weak convergence theorem sharpens and generalizes the existing convergence analysis conducted before, while the strong convergence theorem provides new analysis results on convergence of the online BP training procedure. The results obtained reveal that with any analytic sigmoid activation function, the online BP training procedure is always convergent, which then underlies successful application of the BP neural networks.**

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

I. Introduction

THE backpropagation (BP) neural networks have been widely applied in various areas of scientific research and engineering [1], [2]. It refers to the feedforward neural networks plus the corresponding BP training procedure with which the network weights are updated according to the gradient-descent principle (that is, the gradient method).

There are two practical ways to implement the gradient method: either using the batch scheme or using the online scheme. The batch scheme corresponds to the standard gradient iteration procedure which updates the network weights after all the training examples are processed. Differently, the online scheme is the procedure of updating network weights immediately after one training example is fed. The fed example may be randomly or circularly selected from the given training examples, but should keep periodic in the training set. The online scheme, sometimes called the online gradient method also, has been proven to be more useful and sometimes unique choice in application, especially when the given training examples are huge. Therefore, convergence of the online BP training procedure is a prerequisite of any successful application of BP neural networks.

There have been many convergence analyses of the training procedures of BP neural networks. Convergence of the online BP training procedure in the case when the activation function of the neural networks is linear has been studied in [5], [7], and [8]. For the nonlinear case, a probabilistic asymptotic analysis on convergence of the online BP training procedure as the training examples goes to infinity has been conducted [3], [4], [9]–[11], [14], and [15]. The deterministic convergence analyses of the online BP training procedure were given in [6], [12], [13], and [16]–[23]. The neural networks discussed in [16]–[19], [22], and [23] are, however, two-layered, that is, without hidden elements, and hence, are of very special form. The analysis conducted in [16] is only for classification application and the training examples are supposed to be linearly separable. Moreover, in [13] and [17], the convergence results under the condition that the training examples are linearly independent have been established, which is obviously very restrictive, because the training examples in practice are huge and inevitably linearly dependent. This assumption was then relaxed in [18]–[21]. The results obtained in [18] and [19] are deterministic in nature but the training examples are asked to be fed in a fixed or specific random order. Further results were presented in [20] and [21] for the more general and important cases in which the training examples are allowed to be linearly dependent, and are supplied to a BP neural network with a hidden layer. Nevertheless, the researches in [20] and [21] have all assumed some impractical constraints on setting the step size $\{\eta_m\}$ of the procedure, say, it must satisfy a posterior condition such as

$$\frac{1}{\eta_0}\left(\left\|\sum_{j=0}^{J-1}\nabla E_{j,\omega}(\omega^0,V^0)\right\|^2+\sum_{i=1}^{n}\left\|\sum_{j=0}^{J-1}\nabla E_{j,v_i}(\omega^0,V^0)\right\|^2\right)\geq\sum_{j=0}^{J-1}\left\|\nabla E_{j,\omega}(\omega^0,V^0)\right\|^2+\sum_{i=1}^{n}\sum_{j=0}^{J-1}\left\|\nabla E_{j,v_i}(\omega^0,V^0)\right\|^2 \tag{1}$$

(the meaning of each notation here will be clarified in the next section), which cannot be verified before the BP training starts. Additionally, the convergence results they obtained have only concluded the convergence of the gradient of the error function (namely, $\nabla E(\omega^{mJ+j},V^{mJ+j})\to 0$) and by no means justified the convergence of the online BP training procedure itself (that is, the convergence of weight sequence $\{(\omega^{mJ+j},V^{mJ+j})\}$).

In this paper, we present a generalized deterministic analysis on convergence of the online BP training procedure for the general case when the training examples are allowed to be

Manuscript received October 22, 2007; revised February 20, 2009; accepted June 09, 2009. First published August 18, 2009; current version published October 07, 2009. This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2007CB31002, by the Key Project of National Natural Science Foundation under Grant 70531030, and by the National Natural Science Foundation under Grant 60575045.

Z.-B. Xu and W.-F. Jing are with the Institute for Information and System Science, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: zbxu@mail.xjtu.edu.cn; wfjing@mail.xjtu.edu.cn).

R. Zhang is with the Institute for Information and System Science, Xi'an Jiaotong University, Xi'an 710049, China and also with the Department of Mathematics, Northwest University, Xi'an 710069, China (e-mail: rzhang@nwu.edu.cn).

Digital Object Identifier 10.1109/TNN.2009.2025946

linearly dependent, and the networks are of general form with hidden neurons. We will establish a generic convergence theory of the online BP training procedure that sharpens and generalizes the existing results (particularly, those of [20] and [21]) in the following sense: 1) the convergence of weight sequence $\{(\omega^{mJ+j}, V^{mJ+j})\}$ is concluded; 2) all the posterior assumptions on step size η_m are dismissed; and 3) several general convergence theorems of the online BP training procedure are proven in much weaker conditions than those assumed before.

We will show that the conditions for assuring the convergence of the online BP training will be met with any analytic error function. So the use of any analytic sigmoid function as activation function, as the case of ordinary application, can naturally yield convergence of the online BP training. This underlies successful application of the BP neural networks.

Note that the BP training procedure, as the gradient method in optimization techniques, is the most fundamental method for the training of neural networks. Most other training algorithms such as regularization methods, conjugate gradient method, and Newton methods can be seen as the variants of the BP training procedure in a certain way. Therefore, to clarify convergence of the BP training procedure is the first step towards a full understanding of other more elaborate training algorithms, and perhaps an indispensable part of convergence analysis when a more generic training algorithm is considered. So the results obtained in this paper are of significance not only for the BP training but also for any other training algorithm. In particular, the methodology used in this paper can shed some light on convergence of other (at least gradient-like) training procedures.

The remainder of this paper is organized as follows. In the next section, we formulate mathematically the BP training problem and procedure. We present the main results in Section III with a series of necessary lemmas. The rigorous proofs of the main results and lemmas are presented in Section IV. In Section V, we conclude the paper with some useful remarks.

II. The Online BP Training: A Formulation

Without loss of generality, we consider the three-layer feedforward neural networks with p input, n hidden neurons, and one output neuron. The activation functions used in the hidden and output neurons are all the same, a continuously differentiable function, denoted henceforth by $g : R \longrightarrow R$.

Let $D_J = \{(\xi^j, O^j)\}_{j=0}^{J-1} \subset R^p \times R$ be the given training example set. Denote by

$$V = (v_{ij})_{n\times p} = (v_1, v_2, \ldots, v_n)^T$$
$$v_i = (v_{i1}, v_{i2}, \ldots, v_{ip}) \in R^p, \qquad i = 1, 2, \ldots, n$$

the weight matrix connecting the input and hidden layers of the networks, and

$$\omega = (\omega_1, \omega_2, \ldots, \omega_n)^T \in R^n$$

the weight vector connecting the hidden and output layers. We write

$$G(x) = (g(x_1), g(x_2), \ldots, g(x_n))^T \qquad \forall x \in R^n.$$

Clearly, for any input $\xi \in R^p$, the output of the hidden layer is $G(V\xi - \theta)$ where $\theta = (\theta_1, \theta_2, \ldots, \theta_n)^T \in R^n$ is the threshold in the hidden-layer output. Let $\overline{V} = (V, \theta) \in R^{n\times(p+1)}$, $\overline{\xi} = (\xi, -1) \in R^{p+1}$. Then, $G(V\xi - \theta) = G(\overline{V}\overline{\xi})$. Therefore, without loss of generality, here we can suppose that $\theta = 0$. In the same way, the final output of the networks can be written as

$$\zeta = g(\omega \cdot G(V\xi))$$

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

For any fixed weights (ω, V), the error of the neural networks is defined as

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

where

$$E_j(\omega, V) = \frac{1}{2}\left(O^j - g\left(\omega \cdot G(V\xi^j)\right)\right)^2, \qquad j = 0, 1, \ldots, J-1.$$

The neural network training problem is then to look for the optimal choice (ω^*, V^*) of the weights so as to minimize the error function $E(\omega, V)$.

The BP training scheme is an approach to find (ω^*, V^*) through applying the gradient-descent method, combined with the BP scheme of computation for gradient of the error function [26].

We easily calculate that

$$\begin{aligned}
\nabla E_{j,\omega}(\omega, V) &:= \left(\frac{\partial E_j(\omega, V)}{\partial \omega_1}, \ldots, \frac{\partial E_j(\omega, V)}{\partial \omega_n}\right)^T \\
&= -(O^j - \zeta^j) g'\left(\omega \cdot G(V\xi^j)\right) G(V\xi^j) \\
\nabla E_{j,V}(\omega, V) &:= \left(\nabla E_{j,v_1}(\omega, V), \ldots, \nabla E_{j,v_n}(\omega, V)\right)^T \\
\nabla E_{j,v_i}(\omega, V) &:= \left(\frac{\partial E_j(\omega, V)}{\partial v_{i1}}, \ldots, \frac{\partial E_j(\omega, V)}{\partial v_{ip}}\right) \\
&= -(O^j - \zeta^j) g'\left(\omega \cdot G(V\xi^j)\right) g'(v_i\xi^j)\omega_i\xi^j \\
\nabla E(\omega, V) &= \sum_{j=0}^{J-1} \nabla E_j(\omega, V) \\
&= \left(\sum_{j=0}^{J-1} \nabla E_{j,\omega}(\omega, V), \sum_{j=0}^{J-1} \nabla E_{j,V}(\omega, V)\right) \\
&= \left(\nabla E_\omega(\omega, V), \nabla E_V(\omega, V)\right), \\
&\qquad i = 1, \ldots, n, \qquad j = 0, \ldots, J-1. \qquad (2)
\end{aligned}$$

Then, the online BP training procedure can be formulated as the following iteration procedure:

$$\begin{aligned}
\omega^{mJ+j+1} &= \omega^{mJ+j} - \eta_m \nabla E_{j,\omega}(\omega^{mJ+j}, V^{mJ+j}) \\
v_i^{mJ+j+1} &= v_i^{mJ+j} - \eta_m \nabla E_{j,v_i}(\omega^{mJ+j}, V^{mJ+j}), \\
&\qquad i = 1, 2, \ldots, n, \qquad j = 0, 1, \ldots, J-1, \\
&\qquad m = 0, 1, \ldots \qquad (3)
\end{aligned}$$

Here η_m is the step size or, as it is called, the learning rate, whose value may be changed after each cycle of iteration.

Equation (3) is the formulation of the online BP training procedure that we will focus on in this paper with a detailed analysis.

As in [20] and [21], the analysis of the online BP training procedure will be conducted under a set of assumptions. We formulate our assumptions as follows:

(A1) $|g(t)|$ and $|g'(t)|$ are uniformly bounded for $t \in R$ and $g'(t)$ is Lipschitz continuous, that is, there is a positive constant L such that

$$\left|g'(t) - g'(\bar{t})\right| \le L|t - \bar{t}| \qquad \forall t, \bar{t} \in R;$$

(A2) the weight sequence $\{||(\omega^k, V^k)|| : k = 1, 2, \ldots\}$ is upper bounded;

(A3) $\lim_{m\to\infty}(\eta_m/\eta_{m+1}) = 1$ and $\eta_m \in l^2 \setminus l^1$, that is

$$\sum_{m=1}^{\infty} \eta_m = \infty \quad \text{and} \quad \sum_{m=1}^{\infty} \eta_m^2 < \infty.$$

Remark 1: For comparison purpose, we also list the assumptions used in [20] and [21] as follows:

(W1) $|g(t)|$, $|g'(t)|$ and $|g''(t)|$ are uniformly bounded for $t \in R$;

(W2) the weight sequence $\{||\omega^k||, k = 1, 2, \ldots\}$ is bounded;

(W3) the initial learning rate η_0 satisfies (1);

(W4) $\eta_m = \eta_0/(1 + m\beta\eta_0)$ with $\beta \in (\beta_0, 1/\eta_0)$, where $\beta_0 = 8(1 + (1 + \gamma)(1 + \gamma^{-1})\gamma_1)$ for some positive constants γ and γ_1, $m = 1, 2, \ldots$

Comparing (A1)–(A3) with (W1)–(W4), we see that a weaker condition "$|g'(t)|$ is Lipschitz continuous" in (A1) is used to replace the stronger condition "$|g''(t)|$ is uniformly bounded" in (W1); conditions (A2) and (W2) both include the boundedness of $\{||\omega^k||\}$. This seems sufficient for the weak convergence analysis in [20] and [21], but we must suppose the similar boundedness of $\{||(\omega^k, V^k)||\}$ instead of $\{||\omega^k||\}$ in order to derive a strong convergence result. As pointed out in the Introduction, (W3) is a posterior condition on the step size η_m that cannot be tested offline, and it is dismissed in our assumptions; (W4) essentially implies that $\eta_m = O(1/m)$, which obviously satisfies all the three conditions in (A3). So (W4) is just a special case of (A3). Besides the setting of η_m as in (W4), there are also other possible choices of η_m meeting with (A3), such as $\eta_m = K/m^\alpha$ where K and α are any positive constants with $(1/2) < \alpha \le 1$. Moreover, different from (W4), η_m in (A3) does not depend on the initial learning rate η_0 any more. That is to say, no matter how η_0 is, the convergence of the online BP training procedure will always follow. This relaxes the loading of specifying the initial learning rate η_0. It is clear that (A3) is a more general and weaker setting for the step size η_m than (W4). This shows that our assumptions are generally weaker than those assumed in [20] and [21].

Remark 2: It should be noted that the formulation of the online BP training procedure presented in this section is by no means the most generic form. An infinite number of samples can be fed one by one in the training according to some definite physical processes. It is also possible to feed the training samples according to a random mechanism. All these schemes of training cannot be included in the formulation (3). Nevertheless, the model presented in this section is typical in ordinary application of BP neural networks.

III. Main Results

In this section, we summarize the main results we have obtained for convergence of the online BP training procedure (3) under assumptions (A1)–(A3). The proofs of the results are, however, postponed to the next section so as to make the presentation more readable.

For any $\omega \in R^n$, $V \in R^{n\times p}$, we denote by (ω, V) the matrix $[\omega\vdots V]$, and by $||(\omega, V)||$ the Frobenius norm of $[\omega\vdots V]$, which is defined by

$$||(\omega, V)|| = \left(||\omega||^2 + ||V||_F^2\right)^{\frac{1}{2}} \tag{4}$$

with

$$||V||_F^2 = \sum_{i=1}^{n} ||v_i||^2 = \sum_{i=1}^{n}\sum_{j=1}^{p} v_{ij}^2.$$

Let $\{(\omega^{mJ+j}, V^{mJ+j})\}_{m=1}^{\infty}$ be the weight sequence defined by (3). Then, in the following, we say that the procedure (3) is weakly convergent whenever $\{\nabla E(\omega^{mJ+j}, V^{mJ+j})\}$ converges to zero as $m \to \infty$, and, it is strongly convergent whenever $\{(\omega^{mJ+j}, V^{mJ+j})\}$ converges to a limit (ω^*, V^*) such that $\nabla E(\omega^*, V^*) = 0$.

Note that, in these terms, the weak convergence of (3) implies that the error function E will decrease to a stable value as $m \to \infty$, while the strong convergence implies that the weight sequence will stabilize to a fixed value at which the error function attains its minimum (may be a local minimum).

A. Weak Convergence

We need to establish a series of lemmas in order for the main theorems to be proven.

First, we derive the boundedness and the Lipschitz continuity of $\nabla E(\omega, V)$ from the assumptions (A1) and (A2). We have the following.

Lemma A1: Under assumptions (A1) and (A2):

i) $\|\nabla E(\omega, V)\|$ is bounded;

ii) $\nabla E(\omega, V)$ satisfies the Lipschitz condition in the sense that there exists a positive constant L' such that for any $\theta \in [0, 1]$, $l, k = 0, 1, \ldots$

$$\begin{aligned}\left\|\nabla E_{j,\omega}(\omega^l, V^l) - \nabla E_{j,\omega}(\omega^k, V^k)\right\| &\le L' \left\|(\omega^l, V^l) - (\omega^k, V^k)\right\| \\ \left\|\nabla E_{j,V}(\omega^l, V^l) - \nabla E_{j,V}(\omega^k, V^k)\right\| &\le L' \left\|(\omega^l, V^l) - (\omega^k, V^k)\right\|\end{aligned}$$

and, furthermore

$$\left\|\nabla E\left((\omega^k, V^k) + \theta\left((\omega^l, V^l) - (\omega^k, V^k)\right)\right) - \nabla E(\omega^k, V^k)\right\| \le L'\theta\left\|(\omega^l, V^l) - (\omega^k, V^k)\right\|.$$

The next lemma gives a useful reformulation of the online BP training procedure (3) based on which the convergence analysis of the procedure will be conducted.

Lemma A2: Under assumptions (A1)–(A3), for any $m = 0, 1, \ldots$ and $j = 0, 1, \ldots, J-1$, there holds

$$\begin{aligned}\omega^{(m+1)J+j} &= \omega^{mJ+j} - \eta_m \nabla E_\omega(\omega^{mJ+j}, V^{mJ+j}) \\ &\quad + \gamma_m(\omega^{mJ+j}, V^{mJ+j}) \\ V^{(m+1)J+j} &= V^{mJ+j} - \eta_m \nabla E_V(\omega^{mJ+j}, V^{mJ+j}) \\ &\quad + \gamma_m(\omega^{mJ+j}, V^{mJ+j})\end{aligned}$$

where $\gamma_m(\omega^{mJ+j}, V^{mJ+j})$ obeys the estimation

$$\left\|\gamma_m(\omega^{mJ+j}, V^{mJ+j})\right\| \le C\eta_m^2$$

for a positive scalar C.

From Lemma A2, we can see that the online BP training procedure (3) can be viewed as an ordinary gradient iteration with error [24]. The following lemma is a very useful tool in the analysis of such type of iterations with error.

Lemma A3: Let Y_t, W_t, and Z_t be three sequences such that W_t is nonnegative and Y_t is bounded for all t. If

$$Y_{t+1} \le Y_t - W_t + Z_t, \qquad t = 0, 1, \ldots$$

and the series $\sum_{t=0}^{\infty} Z_t$ is convergent, then Y_t converges to a finite value and $\sum_{t=0}^{\infty} W_t < \infty$.

We also need the following estimation on deviation of error functions $E(\omega^l, V^l)$ and $E(\omega^k, V^k)$.

Lemma A4: For any integers l and k, the following estimation holds:

$$E(\omega^l, V^l) - E(\omega^k, V^k) \le \left\langle \nabla E(\omega^k, V^k), (\omega^l, V^l) - (\omega^k, V^k) \right\rangle + \frac{L'}{2}\left\|(\omega^l, V^l) - (\omega^k, V^k)\right\|^2.$$

Lemma A5: Under assumptions (A1)–(A3), the sequence $\{E(\omega^{mJ+j}, V^{mJ+j})\}$ converges as $m \to \infty$ and

$$\sum_{m=0}^{\infty} \eta_m \left\|\nabla E(\omega^{mJ+j}, V^{mJ+j})\right\|^2 < +\infty.$$

By virtue of Lemmas A1–A5, we can prove the following fundamental weak convergence theorem of the online BP training procedure.

Theorem A: Under assumptions (A1)–(A3), the online BP training procedure (3) is weakly convergent. More precisely, we have

$$\lim_{m\to\infty} \left\|\nabla E(\omega^{mJ+j}, V^{mJ+j})\right\| = 0$$

or, equivalently

$$\lim_{m\to\infty} \left\|\nabla E_{\omega}(\omega^{mJ+j}, V^{mJ+j})\right\| = 0$$
$$\lim_{m\to\infty} \left\|\nabla E_{v_i}(\omega^{mJ+j}, V^{mJ+j})\right\| = 0$$

for all $i = 1, 2, \ldots, n$ and $j = 0, 1, \ldots, J-1$.

B. Strong Convergence

We will establish strong convergence results of the online BP training procedure under some other additional assumptions:

(B1) $E(\omega, V)$ has at most countably infinite number of stationary points;

(B2) $E(\omega, V)$ is directionally convex in the sense that for any ω, ω', v_i, and v_i', there holds

$$\left\langle \nabla E_{\omega}(\omega, V), (\omega - \omega') \right\rangle \ge 0 \tag{5}$$
$$\left\langle \nabla E_{v_i}(\omega, V), (v_i - v_i') \right\rangle \ge 0, \qquad i = 1, 2, \ldots, n; \tag{6}$$

(B3) $\nabla E(\omega, V)$ is a local injection, that is, for any (ω, V), there is a neighborhood Θ of (ω, V) such that whenever $(\omega_1, V_1) \ne (\omega_2, V_2)$ and $(\omega_1, V_1), (\omega_2, V_2) \in \Theta$, there holds $\nabla E(\omega_1, V_1) \ne \nabla E(\omega_2, V_2)$.

We can immediately show that condition (B3) here actually implies condition (B1), so (B3) is just a special case of (B1). To be more precise, we have the following.

Lemma B1: If $E(\omega, V)$ satisfies assumption (B3), then $E(\omega, V)$ contains at most countably infinite number of stationary points.

It should be noted that the directional convexity introduced in (B2) is a much weaker condition than the convexity of $E(\omega, V)$. Actually, as in [25], we easily show that $E(\omega, V)$ is directionally convex in the sense of (5) and (6) if and only if

$$\nabla^2 E_{\omega}(\omega, V) \ge 0 \text{ and } \nabla^2 E_{v_i}(\omega, V) \ge 0, \qquad i = 1, 2, \ldots, n$$

where $\nabla^2 E_{\omega}(\omega, V)$ and $\nabla^2 E_{v_i}(\omega, V)$ denote the Hessian matrices of $E(\omega, V)$ with respect to ω and v_i, respectively. Thus, if we define

$$E(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_i x_j$$

we can directly test that function $E(x_1, x_2, \ldots, x_n)$ is directionally convex (with respect to every direction x_i) if and only if $a_{ii} \ge 0, \forall i = 1, 2, \ldots, n$. Nevertheless, it is well known that $E(x_1, x_2, \ldots, x_n)$ is convex if and only if the matrix $A = (a_{ij})_{n\times n}$ is nonnegative definite. This example highlights the distinction between the directional convexity and the convexity of a function.

Let $\Omega(E) = \{(\omega, V) : \nabla E(\omega, V) = 0\}$ be the stationary point set of the error function $E(\omega, V)$. As the first step of justifying the strong convergence, we need to show that any limit point of the weight sequence $\{(\omega^{mJ+j}, V^{mJ+j})\}$ must be in $\Omega(E)$.

Lemma B2: Under assumptions (A1)–(A3), any limit point (ω^*, V^*) of $\{(\omega^{mJ+j}, V^{mJ+j})\}_{m=0}^{\infty}$ is a stationary point of $E(\omega, V)$, i.e., $(\omega^*, V^*) \in \Omega(E)$.

With Lemmas B1 and B2, we can finally prove the following main result on the strong convergence of the online BP training procedure.

Theorem B: Assume that conditions (A1)–(A3) hold and $E(\omega, V)$ satisfies either (B1), (B2), or (B3). Then, the online BP training procedure (3) is strongly convergent, that is, there is a fixed value $(\omega^*, V^*) \in \Omega(E)$ such that weight sequence $\{(\omega^{mJ+j} V^{mJ+j})\}$ converges to (ω^*, V^*) as $m \to \infty$.

Remark 3: Note that assumptions (B1)–(B3) assumed here are all given in terms of the error function E, which may not be tested beforehand in application. However, we argue that most of the widely used error functions can actually satisfy at least one of the conditions listed. In fact, it is known that any analytic function defined on an open set, if not constant, must contain at most countably infinite number of stationary points. So whenever the error function E is analytic and nonconstant on an open region, it has at most countably infinite stationary points. Through partitioning, we can divide the domain of the error

function E into at most countable open subregions (the partitioning then consists of all the open subregions and their boundaries). The subregions consist of two types: one is the "constant region" on which E takes fixed constant value, and the other is "nonconstant region" which contains no "constant region." It is clear that whenever the iterate of the online BP training reaches a "constant region," the procedure is naturally terminated and arrives at a local minimizer. While the iterate of the online BP training reaches a "nonconstant region," it must stay in that region (at least for sufficiently large m, as it will be proven in Theorem B). So the online BP training procedure must also converge to a local minimizer because E has at most countably infinite number of stationary points on the "nonconstant region." As a result, we can conclude that as long as the activation function g is analytic, the online BP training procedure must be convergent to a local minimizer. It turns out that the use of any analytic sigmoid activation function, like $g(x) = 1/(1+e^{-x})$, as the usual application case, can always assure the strong convergence of the online BP training procedure. However, Theorem B might not imply the strong convergence of the online BP training when the activation function is not analytic, say

$$g(x) = \begin{cases} 1, & x > 1 \\ x, & -1 \le x \le 1 \\ -1, & x < 1. \end{cases}$$

IV. Proofs of Results

The proofs of the lemmas and theorems presented in Section III are given in this section.

Proof of Lemma A1: By formula (2), for any $j = 0, 1, \ldots, J-1$ and $i = 1, \ldots, n$, we have

$$\nabla E_{j,\omega}(\omega, V) = -(O^j - \zeta^j) g'\left(\omega \cdot G(V\xi^j)\right) G(V\xi^j)$$

and

$$\nabla E_{j,v_i}(\omega, V) = -(O^j - \zeta^j) g'\left(\omega \cdot G(V\xi^j)\right) g'(v_i\xi^j)\omega_i\xi^j.$$

Then, by the boundedness of g, g', and ω_i from the assumptions (A1) and (A2), we can easily derive that $\| \nabla E_{j,\omega}(\omega, V)\|$ and $\| \nabla E_{j,v_i}(\omega, V)\|$ are all bounded for every $j = 0, 1, \ldots, J-1$ and $i = 1, \ldots, n$. Therefore, by (2) and (4)

$$\|\nabla E_j(\omega, V)\| = \left(\|\nabla E_{j,\omega}(\omega, V)\|^2 + \sum_{i=1}^{n} \|\nabla E_{j,v_i}(\omega, V)\|^2 \right)^{\frac{1}{2}}$$

and

$$\|\nabla E_\omega(\omega, V)\| \le \sum_{j=0}^{J-1} \|\nabla E_{j,\omega}(\omega, V)\|$$

$$\|\nabla E_V(\omega, V)\| \le \sum_{i=1}^{n} \sum_{j=0}^{J-1} \|\nabla E_{j,v_i}(\omega, V)\|$$

we know that $\| \nabla E_j(\omega, V)\|$, $\| \nabla E_\omega(\omega, V)\|$, and $\| \nabla E_V(\omega, V)\|$ are all bounded for every $j = 0, 1, \ldots, J-1$. This directly implies the boundedness of $\| \nabla E(\omega, V)\| = (\| \nabla E_\omega(\omega, V)\|^2 + \| \nabla E_V(\omega, V)\|^2)^{1/2}$.

Furthermore, since $g'(\omega \cdot G(V\xi^j))$ satisfies the Lipschitz condition and $G(V\xi^j) = (g(v_1\xi^j), g(v_2\xi^j), \ldots, g(v_n\xi^j))^T$ is also Lipschitz continuous by the boundedness of g' on the trajectory $\{(\omega^k, V^k)\}$, we deduce that

$$\begin{aligned}
&\left\|\nabla E_{j,\omega}(\omega^l, V^l) - \nabla E_{j,\omega}(\omega^k, V^k)\right\| \\
&= \left\| \left(O^j - g\left(\omega^l \cdot G(V^l\xi^j)\right)\right) g'\left(\omega^l \cdot G(V^l\xi^j)\right) G(V^l\xi^j) \right. \\
&\qquad \left. -(O^j - g(\omega^k \cdot G(V^k\xi^j)))g'(\omega^k \cdot G(V^k\xi^j))G(V^k\xi^j) \right\| \\
&= \left\| O^j g'\left(\omega^l \cdot G(V^l\xi^j)\right) G(V^l\xi^j) \right. \\
&\qquad - O^j g'\left(\omega^k \cdot G(V^k\xi^j)\right) G(V^k\xi^j) \\
&\qquad - g\left(\omega^l \cdot G(V^l\xi^j)\right) g'\left(\omega^l \cdot G(V^l\xi^j)\right) G(V^l\xi^j) \\
&\qquad \left. + g\left(\omega^k \cdot G(V^k\xi^j)\right) g'\left(\omega^k \cdot G(V^k\xi^j)\right) G(V^k\xi^j) \right\| \\
&\le C_1 \left\| \left(g'(\omega^l \cdot G(V^l\xi^j)\right) - g'(\omega^k \cdot G(V^k\xi^j) \right\| \\
&\quad + C_2 \left\| G(V^l\xi^j) - G(V^k\xi^j) \right\| \\
&\le C_1 L \left\| \omega^l \cdot G(V^l\xi^j) - \omega^k \cdot G(V^k\xi^j) \right\| \\
&\quad + C_2 \left\| G(V^l\xi^j) - G(V^k\xi^j) \right\| \\
&\le C_1' \|\omega^l - \omega^k\| + C_2' \|V^l - V^k\| \\
&\le L_{j,1}' \left(\|\omega^l - \omega^k\| + \|V^l - V^k\| \right) \\
&\le L_{j,1}' \left(2\|\omega^l - \omega^k\|^2 + 2\|V^l - V^k\|^2 \right)^{\frac{1}{2}} \\
&\le L_{j,1} \left\| (\omega^l, V^l) - (\omega^k, V^k) \right\| \qquad (7)
\end{aligned}$$

where C_1, C_2, C_1', C_2', $L_{j,1}'$, and $L_{j,1}$ are appropriate positive constants.

Similarly, we get for any $j = 0, 1, \ldots, J-1$

$$\begin{aligned}
&\left\|\nabla E_{j,v_i}(\omega^l, V^l) - \nabla E_{j,v_i}(\omega^k, V^k)\right\| \\
&\quad = \left\| (O^j - \zeta^j) g'\left(\omega^l \cdot G(V^l\xi^j)\right) g'\left(v_i^l\xi^j\right) \omega_i^l\xi^j \right. \\
&\qquad \left. -(O^j - \zeta^j) g'\left(\omega^k \cdot G(V^k\xi^j)\right) g'\left(v_i^k\xi^j\right) \omega_i^k\xi^j \right\| \\
&\quad \le D_1 \left\| g'\left(\omega^l \cdot G(V^l\xi^j)\right) - g'\left(\omega^k \cdot G(V^k\xi^j)\right) \right\| \\
&\qquad + D_2 \left\| g'\left(v_i^l\xi^j\right) - g'\left(v_i^k\xi^j\right) \right\| \\
&\quad \le D_1' \|\omega^l - \omega^k\| + D_2' \left\| v_i^l - v_i^k \right\|
\end{aligned}$$

so we have

$$\begin{aligned}
&\left\|\nabla E_{j,V}(\omega^l, V^l) - \nabla E_{j,V}(\omega^k, V^k)\right\| \\
&\quad \le \sum_{i=1}^{n} \left\|\nabla E_{j,v_i}(\omega^l, V^l) - \nabla E_{j,v_i}(\omega^k, V^k)\right\| \\
&\quad \le \sum_{i=1}^{n} \left(D_1' \|\omega^l - \omega^k\| + D_2' \left\| v_i^l - v_i^k \right\| \right) \\
&\quad = nD_1' \|\omega^l - \omega^k\| + D_2' \sum_{i=1}^{n} \left\| v_i^l - v_i^k \right\| \\
&\quad \le L_{j,2}' \|\omega^l - \omega^k\| + L_{j,2}'' \|V^l - V^k\| \\
&\quad \le L_{j,2} \left\| (\omega^l, V^l) - (\omega^k, V^k) \right\| \qquad (8)
\end{aligned}$$

where D_1, D_2, D_1', D_2', $L_{j,2}'$, $L_{j,2}''$, and $L_{j,2}$ are all positive constants.

Combining (7) with (8), it thus follows that

$$\begin{aligned}
&\left\|\nabla E\left((\omega^k,V^k)+\theta\left((\omega^l,V^l)-(\omega^k,V^k)\right)\right)-\nabla E(\omega^k,V^k)\right\| \\
&\quad\le\sum_{j=0}^{J-1}\left\|\nabla E_j\left((\omega^k,V^k)+\theta\left((\omega^l,V^l)-(\omega^k,V^k)\right)\right)\right. \\
&\qquad\qquad\left.-\nabla E_j(\omega^k,V^k)\right\| \\
&\quad\le\sum_{j=0}^{J-1}\left(\left\|\nabla E_{j,\omega}\left((\omega^k,V^k)+\theta\left((\omega^l,V^l)-(\omega^k,V^k)\right)\right)\right.\right. \\
&\qquad\qquad\left.-\nabla E_{j,\omega}(\omega^k,V^k)\right\| \\
&\qquad\qquad+\left\|\nabla E_{j,V}\left((\omega^k,V^k)+\theta\left((\omega^l,V^l)-(\omega^k,V^k)\right)\right)\right. \\
&\qquad\qquad\left.\left.-\nabla E_{j,V}(\omega^k,V^k)\right\|\right) \\
&\quad\le\sum_{j=0}^{J-1}\left(L_{j,1}\theta\left\|(\omega^l,V^l)-(\omega^k,V^k)\right\|\right. \\
&\qquad\qquad\left.+L_{j,2}\theta\left\|(\omega^l,V^l)-(\omega^k,V^k)\right\|\right) \\
&\quad\le L_3\theta\left\|(\omega^l,V^l)-(\omega^k,V^k)\right\|\Big\}.
\end{aligned}$$

Let $L'=\max\{L_{j,1},L_{j,2},L_3\}$ and then ii) of Lemma A1 follows.

Proof of Lemma A2: By definition (3), we can write

$$\omega^{(m+1)J}=\omega^{mJ}-\eta_m\sum_{j=0}^{J-1}\bigtriangledown E_{j,\omega}(\omega^{mJ+j},V^{mJ+j})$$

and

$$V^{(m+1)J}=V^{mJ}-\eta_m\sum_{j=0}^{J-1}\bigtriangledown E_{j,V}(\omega^{mJ+j},V^{mJ+j}).$$

So we can express

$$\begin{aligned}
\omega^{(m+1)J}=&\omega^{mJ}-\eta_m\bigtriangledown E_\omega(\omega^{mJ},V^{mJ}) \\
&+\eta_m\sum_{j=0}^{J-1}\left(\bigtriangledown E_{j,\omega}(\omega^{mJ},V^{mJ})\right. \\
&\qquad\left.-\bigtriangledown E_{j,\omega}(\omega^{mJ+j},V^{mJ+j})\right) \\
V^{(m+1)J}=&V^{mJ}-\eta_m\bigtriangledown E_V(\omega^{mJ},V^{mJ}) \\
&+\eta_m\sum_{j=0}^{J-1}\left(\bigtriangledown E_{j,V}(\omega^{mJ},V^{mJ})\right. \\
&\qquad\left.-\bigtriangledown E_{j,V}(\omega^{mJ+j},V^{mJ+j})\right).
\end{aligned}$$

Note that the equation shown at the bottom of the page holds.

$$\begin{aligned}
&\max_j\left\|\left(\omega^{mJ+j},V^{mJ+j}\right)-\left(\omega^{mJ},V^{mJ}\right)\right\| \\
&\quad\le J\eta_m\max_j\left\|\left(\bigtriangledown E_{j,\omega}\left(\omega^{mJ+j},V^{mJ+j}\right),\bigtriangledown E_{j,V}\left(\omega^{mJ+j},V^{mJ+j}\right)\right)\right\| \\
&\quad= J\eta_m\max_j\left\|\left(\bigtriangledown E_{j,\omega}\left(\omega^{mJ+j},V^{mJ+j}\right)-\bigtriangledown E_{j,\omega}\left(\omega^{mJ},V^{mJ}\right)+\bigtriangledown E_{j,\omega}\left(\omega^{mJ},V^{mJ}\right),\right.\right. \\
&\qquad\qquad\left.\left.\bigtriangledown E_{j,V}\left(\omega^{mJ+j},V^{mJ+j}\right)-\bigtriangledown E_{j,V}\left(\omega^{mJ},V^{mJ}\right)+\bigtriangledown E_{j,V}\left(\omega^{mJ},V^{mJ}\right)\right)\right\| \\
&\quad= J\eta_m\max_j\left\{\left\|\bigtriangledown E_{j,\omega}\left(\omega^{mJ+j},V^{mJ+j}\right)-\bigtriangledown E_{j,\omega}\left(\omega^{mJ},V^{mJ}\right)+\bigtriangledown E_{j,\omega}\left(\omega^{mJ},V^{mJ}\right)\right\|^2\right. \\
&\qquad\qquad\left.+\left\|\bigtriangledown E_{j,V}\left(\omega^{mJ+j},V^{mJ+j}\right)-\bigtriangledown E_{j,V}\left(\omega^{mJ},V^{mJ}\right)+\bigtriangledown E_{j,V}\left(\omega^{mJ},V^{mJ}\right)\right\|^2\right\}^{\frac{1}{2}} \\
&\quad\le J\eta_m\max_j\left\{\left[\left\|\bigtriangledown E_{j,\omega}\left(\omega^{mJ+j},V^{mJ+j}\right)-\bigtriangledown E_{j,\omega}\left(\omega^{mJ},V^{mJ}\right)\right\|+\left\|\bigtriangledown E_{j,\omega}\left(\omega^{mJ},V^{mJ}\right)\right\|\right]^2\right. \\
&\qquad\qquad\left.+\left[\left\|\bigtriangledown E_{j,V}\left(\omega^{mJ+j},V^{mJ+j}\right)-\bigtriangledown E_{j,V}\left(\omega^{mJ},V^{mJ}\right)\right\|+\left\|\bigtriangledown E_{j,V}\left(\omega^{mJ},V^{mJ}\right)\right\|\right]^2\right\}^{\frac{1}{2}} \\
&\quad\le J\eta_m\left\{2\max_j\left\|\bigtriangledown E_{j,\omega}\left(\omega^{mJ+j},V^{mJ+j}\right)-\bigtriangledown E_{j,\omega}\left(\omega^{mJ},V^{mJ}\right)\right\|^2\right. \\
&\qquad\qquad+2\max_j\left\|\bigtriangledown E_{j,V}\left(\omega^{mJ+j},V^{mJ+j}\right)-\bigtriangledown E_{j,V}\left(\omega^{mJ},V^{mJ}\right)\right\|^2 \\
&\qquad\qquad\left.+2\max_j\left[\left\|\bigtriangledown E_{j,\omega}\left(\omega^{mJ},V^{mJ}\right)\right\|^2+\left\|\bigtriangledown E_{j,V}\left(\omega^{mJ},V^{mJ}\right)\right\|\right]^2\right\}^{1/2} \\
&\quad\le J\eta_m\left\{2L'^2\max_j\left\|(\omega^{mJ+j},V^{mJ+j})-(\omega^{mJ},V^{mJ})\right\|^2+2L'^2\max_j\left\|(\omega^{mJ+j},V^{mJ+j})-(\omega^{mJ},V^{mJ})\right\|^2\right. \\
&\qquad\qquad\left.+2\max_j\left[\left\|\bigtriangledown E_{j,\omega}(\omega^{mJ},V^{mJ})\right\|^2+\left\|\bigtriangledown E_{j,V}(\omega^{mJ},V^{mJ})\right\|\right]^2\right\}^{1/2} \\
&\quad\le J\eta_m\left\{2L'\max_j\left\|\left(\omega^{mJ+j},V^{mJ+j}\right)-\left(\omega^{mJ},V^{mJ}\right)\right\|+\sqrt{2}\max_j\left[\left\|\bigtriangledown E_{j,\omega}\left(\omega^{mJ},V^{mJ}\right)\right\|^2+\left\|\bigtriangledown E_{j,V}\left(\omega^{mJ},V^{mJ}\right)\right\|^2\right]^{\frac{1}{2}}\right\} \\
&\quad= 2JL'\eta_m\max_j\left\|\left(\omega^{mJ+j},V^{mJ+j}\right)-\left(\omega^{mJ},V^{mJ}\right)\right\|+\sqrt{2}J\eta_m\max_j\left\|\bigtriangledown E_j\left(\omega^{mJ},V^{mJ}\right)\right\|.
\end{aligned}$$

So from the boundedness of $\|\bigtriangledown E_j(\omega^{mJ}, V^{mJ})\|$, we can get

$$\begin{aligned}\max_j \left\|(\omega^{mJ+j}, V^{mJ+j}) - (\omega^{mJ}, V^{mJ})\right\| &\le \frac{\sqrt{2}J\eta_m}{1-2JL'\eta_m} \max_j \left\|\bigtriangledown E_j(\omega^{mJ}, V^{mJ})\right\| \\ &\le \frac{\sqrt{2}JC_3\eta_m}{1-2JL'\eta_m}\end{aligned}$$

where C_3 is a positive constant, and from Lemma A1, we have

$$\begin{aligned}&\left\|\bigtriangledown E_{j,\omega}(\omega^{mJ}, V^{mJ}) - \bigtriangledown E_{j,\omega}(\omega^{mJ+j}, V^{mJ+j})\right\| \\ &\quad\le L' \left\|(\omega^{mJ}, V^{mJ}) - (\omega^{mJ+j}, V^{mJ+j})\right\| \\ &\quad\le \frac{\sqrt{2}JL'C_3\eta_m}{1-2JL'\eta_m}\end{aligned}$$

and

$$\left\|\bigtriangledown E_{j,V}(\omega^{mJ}, V^{mJ}) - \bigtriangledown E_{j,V}(\omega^{mJ+j}, V^{mJ+j})\right\| \le \frac{\sqrt{2}JL'C_3\eta_m}{1-2JL'\eta_m}.$$

So, denoting by $\gamma_m(\omega^{mJ}, V^{mJ})$ the quantity

$$\eta_m \sum_{j=0}^{J-1} \left(\bigtriangledown E_{j,\omega}(\omega^{mJ}, V^{mJ}) - \bigtriangledown E_{j,\omega}(\omega^{mJ+j}, V^{mJ+j})\right)$$

or

$$\eta_m \sum_{j=0}^{J-1} \left(\bigtriangledown E_{j,V}(\omega^{mJ}, V^{mJ}) - \bigtriangledown E_{j,V}(\omega^{mJ+j}, V^{mJ+j})\right)$$

we then conclude that there is a positive constant C such that

$$\left\|\gamma_m(\omega^{mJ}, V^{mJ})\right\| \le C\eta_m^2.$$

By the assumption that $\lim_{m\to\infty}(\eta_m/\eta_{m+1}) = 1$, we can similarly verify that for any $j = 0, 1, \ldots, J-1$, the following expression holds:

$$\omega^{(m+1)J+j} = \omega^{mJ+j} - \eta_m \nabla E_\omega(\omega^{mJ+j}, V^{mJ+j}) + \gamma_m(\omega^{mJ+j}, V^{mJ+j})$$

and

$$V^{(m+1)J+j} = V^{mJ+j} - \eta_m \nabla E_V(\omega^{mJ+j}, V^{mJ+j}) + \gamma_m(\omega^{mJ+j}, V^{mJ+j})$$

where $\|\gamma_m(\omega^{mJ}, V^{mJ})\| \le C\eta_m^2$ for a positive constant C. This implies Lemma A2.

Proof of Lemma A3: It follows directly from [24].

Proof of Lemma A4: Denote

$$g(\theta) = E\left((\omega^k, V^k) + \theta\left((\omega^l, V^l) - (\omega^k, V^k)\right)\right).$$

Then, we have

$$\begin{aligned}&E(\omega^l, V^l) - E(\omega^k, V^k) \\ &\quad= g(1) - g(0) \\ &\quad= \int_0^1 \frac{dg(\theta)}{d\theta} d\theta \\ &\quad= \int_0^1 \left((\omega^l, V^l) - (\omega^k, V^k)\right)^T \\ &\qquad\times \nabla E\left((\omega^k, V^k) + \theta\left((\omega^l, V^l) - (\omega^k, V^k)\right)\right) d\theta \\ &\quad= \left((\omega^l, V^l) - (\omega^k, V^k)\right)^T \nabla E(\omega^k, V^k) \\ &\qquad+ \int_0^1 \left((\omega^l, V^l) - (\omega^k, V^k)\right)^T \\ &\qquad\times \left(\nabla E\left((\omega^k, V^k) + \theta\left((\omega^l, V^l) - (\omega^k, V^k)\right)\right)\right. \\ &\qquad\left. -\nabla E(\omega^k, V^k)\right) d\theta \\ &\quad\le \left\langle \nabla E(\omega^k, V^k), (\omega^l, V^l) - (\omega^k, V^k)\right\rangle \\ &\qquad+ L' \left\|(\omega^l, V^l) - (\omega^k, V^k)\right\|^2 \int_0^1 \theta d\theta \\ &\quad\le \left\langle \nabla E(\omega^k, V^k), (\omega^l, V^l) - (\omega^k, V^k)\right\rangle \\ &\qquad+ \frac{L'}{2} \left\|(\omega^l, V^l) - (\omega^k, V^k)\right\|^2.\end{aligned}$$

This arrives at Lemma 4.

Proof of Lemma A5: According to Lemma A-4, we have

$$\begin{aligned}&E\left(\omega^{(m+1)J+j}, V^{(m+1)J+j}\right) - E(\omega^{mJ+j}, V^{mJ+j}) \\ &\quad\le \left\langle \nabla E_\omega(\omega^{mJ+j}, V^{mJ+j}), \left(\omega^{(m+1)J+j} - \omega^{mJ+j}\right)\right\rangle \\ &\qquad+ \left\langle \nabla E_V(\omega^{mJ+j}, V^{mJ+j}), \left(V^{(m+1)J+j} - V^{mJ+j}\right)\right\rangle \\ &\qquad+ \frac{L'}{2} \left\|\left(\omega^{(m+1)J+j}, V^{(m+1)J+j}\right) - (\omega^{mJ+j}, V^{mJ+j})\right\|^2.\end{aligned}$$

From Lemmas A1 and A2, it is direct to conclude that there exists a positive constant C_4 such that

$$\begin{aligned}&E\left(\omega^{(m+1)J+j}, V^{(m+1)J+j}\right) - E(\omega^{mJ+j}, V^{mJ+j}) \\ &\quad\le -\eta_m \left(\left\|\nabla E_\omega(\omega^{mJ+j}, V^{mJ+j})\right\|^2 \right. \\ &\qquad\left. + \left\|\nabla E_V(\omega^{mJ+j}, V^{mJ+j})\right\|^2\right) + C_4\eta_m^2 \\ &\quad= -\eta_m \left\|\nabla E(\omega^{mJ+j}, V^{mJ+j})\right\|^2 + C_4\eta_m^2 \qquad (9)\end{aligned}$$

So, by using Lemma A-3, we conclude that

$$\sum_{m=0}^{\infty} \eta_m \left\|\bigtriangledown E(\omega^{mJ+j}, V^{mJ+j})\right\|^2 < +\infty$$

and $\{E(\omega^{mJ+j}, V^{mJ+j})\}$ converges as $m \to \infty$. This implies Lemma A5.

Proof of Theorem A: To show that the limit $\lim_{m\to\infty} \| \bigtriangledown E(\omega^{mJ+j}, V^{mJ+j})\| = 0$, we assume the contrary, namely, that

$\lim_{m\to\infty}\| \bigtriangledown E(\omega^{mJ+j}, V^{mJ+j})\| \neq 0$. Then, there is a positive constant $\delta > 0$ such that

$$\limsup_{m\to\infty} \left\|\bigtriangledown E(\omega^{mJ+j}, V^{mJ+j})\right\| = \delta > 0.$$

Note that Lemma A5 shows $\liminf_{m\to\infty}\| \bigtriangledown E(\omega^{mJ+j}, V^{mJ+j})\| = 0$, and hence, for any ε with $0 < \varepsilon < \delta$, $\| \bigtriangledown E(\omega^{mJ+j}, V^{mJ+j})\| < (\varepsilon/2)$ for infinitely many m and also $\| \bigtriangledown E(\omega^{mJ+j}, V^{mJ+j})\| > \varepsilon$ for infinitely many m. Consequently, there is an infinite subset of integers $\Gamma = \{m : \| \bigtriangledown E(\omega^{mJ+j}, V^{mJ+j})\| < (\varepsilon/2)\}$ such that for each $m \in \Gamma$, there exists an integer $i(m) > m$ such that

$$\left\|\bigtriangledown E\left(\omega^{i(m)J+j}, V^{i(m)J+j}\right)\right\| > \varepsilon$$

and for any $m < i < i(m)$

$$\frac{\varepsilon}{2} < \left\|\bigtriangledown E(\omega^{iJ+j}, V^{iJ+j})\right\| < \varepsilon.$$

Thus, we deduce that for all $m \in \Gamma$

$$\begin{aligned}
\frac{\varepsilon}{2} &\le \left\|\bigtriangledown E\left(\omega^{i(m)J+j}, V^{i(m)J+j}\right)\right\| - \left\|\bigtriangledown E(\omega^{mJ+j}, V^{mJ+j})\right\| \\
&\le \left\|\bigtriangledown E\left(\omega^{i(m)J+j}, V^{i(m)J+j}\right) - \bigtriangledown E(\omega^{mJ+j}, V^{mJ+j})\right\| \\
&\le L'\left\|\left(\omega^{i(m)J+j}, V^{i(m)J+j}\right) - (\omega^{mJ+j}, V^{mJ+j})\right\| \\
&= L'\left\|\left(\omega^{i(m)J+j} - \omega^{mJ+j}, V^{i(m)J+j} - V^{mJ+j}\right)\right\| \\
&\le L'\sum_{i=m}^{i(m)-1}\left\|\left(\omega^{(i+1)J+j} - \omega^{iJ+j}, V^{(i+1)J+j} - V^{iJ+j}\right)\right\| \\
&= L'\sum_{i=m}^{i(m)-1}\left\|\left(-\eta_i \bigtriangledown E_\omega(\omega^{iJ+j}, V^{iJ+j}) + \gamma_i(\omega^{iJ+j}, V^{iJ+j}),\right.\right. \\
&\qquad \left.\left. -\eta_i \bigtriangledown E_V(\omega^{iJ+j}, V^{iJ+j}) + \gamma_i(\omega^{iJ+j}, V^{iJ+j})\right)\right\| \\
&= L'\sum_{i=m}^{i(m)-1}\left\{\left\|-\eta_i \bigtriangledown E_\omega(\omega^{iJ+j}, V^{iJ+j}) + \gamma_i(\omega^{iJ+j}, V^{iJ+j})\right\|^2 \right. \\
&\qquad \left. + \left\|-\eta_i \bigtriangledown E_V(\omega^{iJ+j}, V^{iJ+j}) + \gamma_i(\omega^{iJ+j}, V^{iJ+j})\right\|^2\right\}^{\frac{1}{2}} \\
&\le L'\sum_{i=m}^{i(m)-1}\left\{2\eta_i^2\left[\left\|\bigtriangledown E_\omega(\omega^{iJ+j}, V^{iJ+j})\right\|^2 + \left\|\bigtriangledown E_V(\omega^{iJ+j}, V^{iJ+j})\right\|^2\right] + 4C^2\eta_i^4\right\}^{\frac{1}{2}} \\
&\le \sqrt{2}L'\sum_{i=m}^{i(m)-1}\eta_i\left\|\bigtriangledown E(\omega^{iJ+j}, V^{iJ+j})\right\| + 2CL'\sum_{i=m}^{i(m)-1}\eta_i^2 \\
&\le \sqrt{2}L'\varepsilon\sum_{i=m}^{i(m)-1}\eta_i + 2CL'\sum_{i=m}^{i(m)-1}\eta_i^2.
\end{aligned}$$

Since $\sum_{m=1}^{\infty}\eta_m^2$ converges by Assumption (A3), it then follows that

$$\liminf_{m\to\infty}\sum_{i=m}^{i(m)-1}\eta_i \ge \frac{1}{2\sqrt{2}L'}. \tag{10}$$

On the other hand, from Lemmas A1 and A2, we have

$$\begin{aligned}
&\left\|\bigtriangledown E\left(\omega^{(m+1)J+j}, V^{(m+1)J+j}\right)\right\| - \left\|\bigtriangledown E(\omega^{mJ+j}, V^{mJ+j})\right\| \\
&\quad\le \left\|\bigtriangledown E\left(\omega^{(m+1)J+j}, V^{(m+1)J+j}\right) - \bigtriangledown E(\omega^{mJ+j}, V^{mJ+j})\right\| \\
&\quad\le L'\left\|\left(\omega^{(m+1)J+j}, V^{(m+1)J+j}\right) - (\omega^{mJ+j}, V^{mJ+j})\right\| \\
&\quad\le \sqrt{2}L'\eta_m\left\|\nabla E(\omega^{mJ+j}, V^{mJ+j})\right\| + 2CL'\eta_m^2
\end{aligned}$$

which implies that for all $m \in \Gamma$, as long as they are sufficiently large so that $L'\eta_m < (\varepsilon/4\sqrt{2})$ and $CL'\eta_m^2 < (\varepsilon/8)$, we have

$$\left\|\bigtriangledown E(\omega^{mJ+j}, V^{mJ+j})\right\| \ge \frac{\varepsilon}{4}.$$

According to (9), we thus have

$$\begin{aligned}
E\left(\omega^{i(m)J+j}, V^{i(m)J+j}\right) &\le E(\omega^{mJ+j}, V^{mJ+j}) \\
&\quad - \left(\frac{\varepsilon}{4}\right)^2\sum_{i=m}^{i(m)-1}\eta_i + C_4\sum_{i=m}^{i(m)-1}\eta_i^2
\end{aligned}$$

for any $m \in \Gamma$.

By the convergence of $E(\omega^{mJ+j}, V^{mJ+j})$ already shown in Lemma A5 and by the assumption $\sum_{m=0}^{\infty}\eta_m^2 < +\infty$, this then implies

$$\liminf_{m\to\infty}\sum_{i=m}^{i(m)-1}\eta_i = 0 < \frac{1}{2\sqrt{2}L'}.$$

This contradicts to (10). The contradiction shows that

$$\lim_{m\to\infty}\left\|\bigtriangledown E(\omega^{mJ+j}, V^{mJ+j})\right\| = 0$$

and, in particular

$$\lim_{m\to\infty}\left\|\nabla E_\omega(\omega^{mJ+j}, V^{mJ+j})\right\| = 0$$

and

$$\lim_{m\to\infty}\left\|\nabla E_\omega(\omega^{mJ+j}, V^{mJ+j})\right\| = 0.$$

The proof of Theorem A is thus completed.

Proof of Lemma B1: Denote $F(x) = \bigtriangledown E(\omega, V)$ where $x = (\omega, V)$, and let $F^{-1}(0)$ be the set of stationary points of $F(x)$ (that is, every element x in $F^{-1}(0)$ satisfies $F(x) = 0$). For any integer n, let B_n be the ball centered at zero with radius n (i.e., $B_n = \{x : \|x\| \le n\}$) and $\Omega_n = B_n \cap \Omega$. Then, it is clear that $\Omega = \bigcup_{n=1}^{\infty}\Omega_n$ and each Ω_n is a compact set. Since, by assumption, F is locally injective, for each $x \in \Omega_n$, there is a neighborhood $\Theta(x)$ such that F is injective when restricted to $\Theta(x)$. The family of the neighborhoods $\{\Theta(x) : x \in \Omega_n\}$

clearly constitutes a compact cover of the set Ω_n. The compactness of Ω_n then implies that there is a finite number of neighborhoods, say, $\{\Theta(x_{ni}) : i = 1, 2, \dots k_n\}$, such that they still cover the set Ω_n. Therefore, $\Omega_n = \bigcup_{i=1}^{k_n} \Theta(x_{ni})$ and, furthermore, we can express

$$\Omega = \bigcup_{n=1}^{\infty} \bigcup_{i=1}^{k_n} \Theta(x_{ni}).$$

In each $\Theta(x_{ni})$, $F(x) = 0$ clearly has at most one solution since, otherwise, two solutions x_1, x_2 with $x_1 \neq x_2$ exist, which could lead to an obvious contradiction: $0 = F(x_1) \neq F(x_2) = 0$. In light of the countable decomposition of Ω above, $F^{-1}(0)$ thus contains at most countably infinite number of elements. That is, $E(\omega, V)$ contains at most countably infinite number of stationary points, justifying Lemma B1.

Proof of Lemma B2: Without loss of generality, we assume that $(\omega^{m_i J+j}, V^{m_i J+j})$ converges to (ω^*, V^*) for a subsequence $\{m_i\}$ of $\{m\}$. Then, by continuity of $\| \bigtriangledown E(\omega, V)\|$ and Theorem A, we have

$$\begin{aligned}\|\bigtriangledown E(\omega^*, V^*)\| &= \lim_{i\to\infty} \left\|\bigtriangledown E(\omega^{m_i J+j}, V^{m_i J+j})\right\| \\ &= \lim_{m\to\infty} \left\|\bigtriangledown E(\omega^{mJ+j}, V^{mJ+j})\right\| = 0\end{aligned}$$

which shows that (ω^*, V^*) is a stationary point of E.

Proof of Theorem B: According to Lemma B1, we can prove the theorem under either assumption (B1) or (B2).

Let us first assume that (B1) holds. Denoting by W the limit set of sequence $\{(\omega^k, V^k)\}$, then assumption (A2) and Lemma B2 show that W is nonempty and all elements of W are stationary points of E. Assume that there are two elements, say (ω^*, V^*) and (ω', V') in W such that $(\omega^*, V^*) \neq (\omega', V')$, $(\omega^{t_l}, V^{t_l}) \to (\omega^*, V^*)$, and $(\omega^{s_l}, V^{s_l}) \to (\omega', V')$ for two subsequences $\{t_l\}$ and $\{s_l\}$ in $(0, \infty)$. Write

$$\begin{aligned}\omega^k &= (\omega_1(k), \omega_2(k), \dots, \omega_n(k)) \\ V^k &= (v_{11}(k), \dots, v_{1p}(k), \dots, v_{n1}(k), \dots, v_{np}(k)) \\ \omega^* &= (\omega_1^*, \omega_2^*, \dots, \omega_n^*) \\ V^* &= \left(v_{11}^*, , \dots, v_{1p}^*, v_{21}^*, \dots, v_{2p}^*, \dots, v_{n1}^*, \dots, v_{np}^*\right) \\ \omega' &= (\omega_1', \omega_2', \dots, \omega_n') \\ V' &= \left(v_{11}', , \dots, v_{1p}', v_{21}', \dots, v_{2p}', \dots, v_{n1}', \dots, v_{np}'\right).\end{aligned}$$

This then particularly implies that

$$\begin{aligned}\omega_i(t_l) &\to \omega_i^* \\ v_{ij}(t_l) &\to v_{ij}^* \\ \omega_i(s_l) &\to \omega_i' \\ v_{ij}(s_l) &\to v_{ij}'\end{aligned}$$

as $l \to \infty$ for all $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, p$. For any fixed real number $\lambda \in (0, 1)$, let $\omega_1^\lambda = \lambda\omega_1^* + (1-\lambda)\omega_1'$. Since we clearly have $(\omega^{k+1}, V^{k+1}) - (\omega^k, V^k) \to 0$ as $k \to \infty$, ω_1^λ must be a limit point of $\{\omega_1(k)\}$ also, thus there is subsequence $(\omega^{r_{k_1}}, V^{r_{k_1}})$ that satisfies $\omega_1(r_{k_1}) \to \omega_1^\lambda$ as $k_1 \to \infty$. Let us consider the sequence $\{\omega_2(r_{k_1})\}$. It is bounded clearly and hence, has a convergent subsequence, say, $\{\omega_2(r_{k_2})\} \subset \{\omega_2(r_{k_1})\}$. Thus, we can define $\omega_2^\lambda = \lim_{k_2\to\infty} \omega_2(r_{k_2})$. Similarly, through considering the sequence $\{\omega_i(r_{k_{i-1}})\}$, we can find a convergent subsequence $\{\omega_i(r_{k_i})\} \subset \{\omega_i(r_{k_{i-1}})\}$ and define $\omega_i^\lambda = \lim_{k_i\to\infty} \omega_i(r_{k_i})$. Continuing this procedure up to n steps, then we can get a decreasing subsequences $\{r_{k_1}\} \supset \{r_{k_2}\} \supset \cdots \supset \{r_{k_n}\}$ with $\omega_i^\lambda = \lim_{k_i\to\infty} \omega_i(r_{k_i})$ for all $i = 1, 2, \dots, n$. Similarly, in doing so, we can also find a decreasing subsequences $\{r_{k_1}\} \supset \{r_{k_2}\} \supset \cdots \supset \{r_{k_{np}}\}$ so that $v_{ij}^\lambda = \lim_{k_{(i-1)p+j}\to\infty} v_{ij}(r_{k_{(i-1)p+j}})$ is well defined. Let $(\omega^\lambda, V^\lambda) = ((\omega_1^\lambda, \omega_2^\lambda, \dots, \omega_n^\lambda), (v_{11}^\lambda, v_{12}^\lambda, \dots, v_{np}^\lambda))$. This then shows that $(\omega^\lambda, V^\lambda)$, for any $\lambda \in (0, 1)$, is a limit point of $\{(\omega^k, V^k)\}$, that is, $(\omega^\lambda, V^\lambda) \in W$. Thus, W contains uncountably infinite number of elements because $\{(\omega^\lambda, V^\lambda)\}$ does. However, by the assumption, this is impossible. So W must contain only one element (say, (ω^*, V^*)). Because of the boundedness of $\{(\omega^k, V^k)\}$, (ω^*, V^*) must be the unique limit of the trajectory $\{(\omega^k, V^k)\}$ (i.e., $(\omega^k, V^k) \to (\omega^*, V^*)$ as $k \to \infty$). This verifies the strong convergence of $\{(\omega^k, V^k)\}$.

Next, let us assume (B2) holds. In this case, for any stationary point (ω^*, V^*) of $E(\omega, V)$, according to Lemma A2, we have

$$\begin{aligned}&\left\|\omega^{(m+1)J+j} - \omega^*\right\|^2 \\ &\quad= \left\|\omega^{mJ+j} - \omega^* - \eta_m \nabla E_\omega(\omega^{mJ+j}, V^{mJ+j})\right. \\ &\qquad\left. +\gamma_m(\omega^{mJ+j}, V^{mJ+j})\right\|^2 \\ &\quad\le \|\omega^{mJ+j} - \omega^*\|^2 \\ &\qquad - 2\eta_m \left\langle \nabla E_\omega(\omega^{mJ+j}, V^{mJ+j}), \omega^{mJ+j} - \omega^*\right\rangle + C_5\eta_m^2\end{aligned}$$

for a positive constant C_5. Also, by Assumption (B2), $\langle \nabla E_\omega(\omega^{mJ+j}, V^{mJ+j}), \omega^{mJ+j} - \omega^*\rangle \ge 0$. So we deduce that

$$\left\|\omega^{(m+1)J+j} - \omega^*\right\|^2 \le \|\omega^{mJ+j} - \omega^*\|^2 + C_5\eta_m^2.$$

According to Assumption (A2) and Lemma A3, we have that for any $j = 0, 1, \dots, J-1$, the limit $\lim_{m\to\infty} \|\omega^{mJ+j} - \omega^*\|$ exists, so does the limit $\lim_{m\to\infty} \|V^{mJ+j} - V^*\|$, that is, the limit

$$\lim_{m\to\infty} \left\|(\omega^{mJ+j}, V^{mJ+j}) - (\omega^*, V^*)\right\| \tag{11}$$

exists.

On the other hand, since $\{(\omega^{mJ+j}, V^{mJ+j})\}$ is bounded, there is a limit point (ω_*, V_*) such that

$$\lim_{i\to\infty} \left\|(\omega^{m_i J+j}, V^{m_i J+j}) - (\omega_*, V_*)\right\| = 0 \tag{12}$$

where $\{(\omega^{m_i J+j}, V^{m_i J+j})\}$ is a subsequence of $\{(\omega^{mJ+j}, V^{mJ+j})\}$ and, by Lemma B2, (ω_*, V_*) is also a stationary state of $E(\omega, V)$. Thus, we can use (ω_*, V_*) in place of (ω^*, V^*) in (11) to deduce the existence of limit $\lim_{m\to\infty} \|(\omega^{mJ+j}, V^{mJ+j}) - (\omega_*, V_*)\|$. Combined with (12), we then conclude that $\lim_{m\to\infty} \|(\omega^{mJ+j}, V^{mJ+j}) - (\omega_*, V_*)\| = 0$, that is, the sequence $\{(\omega^{mJ+j}, V^{mJ+j})\}$ converges to (ω_*, V_*).

The proof of Theorem B is then completed.

V. Conclusion

In this paper, we have analyzed the deterministic convergence of the online BP training procedure for one-hidden-layer

backpropagation neural networks. Two general theorems have been proven, with one claiming the convergence of the gradient sequence $\{\nabla E(\omega^{mJ+j}, V^{mJ+j})\}$ of the error function (the weak convergence), and the other concluding the convergence of weight sequence $\{(\omega^{mJ+j}, V^{mJ+j})\}$ (the strong convergence) under mild conditions. While the strong convergence result is new, the weak convergence theorem sharpens and generalizes those existing analyses (particularly, the results obtained recently by Wu *et al.* [20] and Li *et al.* [21]) in the sense that it is validated not only for much general types of neural networks [say, relaxed assumption (W1) to assumption (A1)], but also for a very general family of learning rates. Different from those proven in [20] and [21], our analysis dismissed the posterior condition (W3) on the step size η_0 and relaxed condition (W4) into the more general and weaker condition (A3). On the one hand, our convergence results have nothing to do with the initial value, and, on the other hand, we have provided wider selections of the step size η_m, which may be very helpful in application. The obtained results settle down the long-standing problem on convergence of the online BP training procedure for the BP neural networks with hidden layers. It is concluded that with any analytic sigmoid activation function, the online BP training procedure is always convergent, which then underlies successful application of the BP neural networks.

Open problems remaining for further research include: uncovering the necessary and sufficient (or the weakest) condition for strong convergence of the online BP training procedure; proving strong convergence of the procedure under some other mild conditions; studying convergence of the online BP training procedure in other step-size setting; and comparing the convergence speed of using different step-size rules.

REFERENCES

[1] Y. C. Liang, W. Z. Lin, H. P. Lee, S. P. Lim, K. H. Lee, and H. Sun, "Proper orthogonal decomposition and its application—Part II: Model reduction for MEMS dynamical analysis," *J. Sound Vib.*, vol. 256, pp. 515–532, 2002.

[2] C. G. Looney, *Pattern Recognition Using Neural Networks*. New York: Oxford Univ. Press, 1997.

[3] T. L. Fine and S. Mukherjee, "Parameter convergence and learning curves for neural networks," *Neural Comput.*, vol. 11, pp. 747–769, 1999.

[4] 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.

[5] Z. Luo, "On the convergence of the LMS algorithm with adaptive learning rate for linear feedforward networks," *Neural Comput.*, vol. 3, no. 2, pp. 226–245, 1991.

[6] Z. 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.

[7] P. Sollich and D. Barber, "Online learning from finite training sets and robustness to input bias," *Neural Comput.*, vol. 10, no. 8, pp. 2201–2217, 1998.

[8] D. P. Bertsekas, *Nonlinear Programming*. Boston, MA: Athena Scientific, 1995.

[9] S. H. Oh, "Improving the error bp algorithm with a modified error function, IEEE Trans. Neural Networks," *IEEE Trans. Circuits Syst.*, vol. 8, no. 3, pp. 799–803, May 1997.

[10] 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.

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

[12] O. L. Manggasarian and M. V. Solodov, "Serial and parallel backpropagation convergence via nonmonotone perturbed minimization," *Optim. Methods Softw.*, vol. 4, pp. 103–116, 1994.

[13] 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. Explosion*, vol. 23, no. 1, pp. 83–97, 2003.

[14] H. J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications*. Berlin, Germany: Springer-Verlag, 1997.

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

[16] W. Wu and Z. Shao, "convergence of an online gradient methods for continuous perceptrons with linearly separable training patterns," *Appl. Math. Lett.*, vol. 16, no. 2, pp. 999–1002, 2003.

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

[18] 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.

[19] Z. Li, W. Wu, and Y. Tian, "Convergence of an online gradient method for FNN with stochastic inputs," *J. Comput. Appl. Math.*, vol. 163, pp. 165–176, 2004.

[20] 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.

[21] Z. X. Li, W. Wu, G. R. Feng, and H. Lu, "Convergence of an online gradient method for BP neural networks with stochastic inputs," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, 2005, vol. 3601, pp. 720–729.

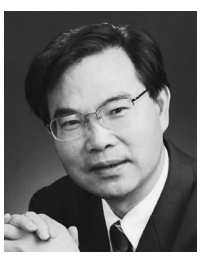
[22] W. Wu, H. Shao, and D. Qu, "Strong convergence for gradient methods for BP networks training," in *Proc. Int. Conf. Neural Netw. Brains*, 2005, pp. 332–334.

[23] N. Zhang, W. Wu, and G. Zheng, "Convergence of gradient method with momentum for two-layer feedforward neural networks with stochastic inputs," *IEEE Trans. Neural Netw.*, vol. 17, no. 2, pp. 522–525, Mar. 2006.

[24] D. P. Bertsekas and J. N. Tsitsiklis, "Gradient convergence in gradient methods with errors," *SIAM J. Optim.*, vol. 10, no. 3, pp. 627–642, 2000.

[25] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations With Several Variables*. New York: Academic, 1970.

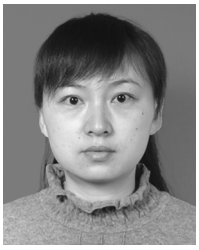
[26] S. Haykin, *Neural Networks: A Comprehensive Foundation*. New York: Macmillan, 1994.



Zong-Ben Xu received the M.S. and Ph.D. degrees in mathematics from Xi'an Jiaotong University, Xi'an, China, in 1981 and 1987, respectively.

In 1988, he was a Postdoctoral Researcher in the Department of Mathematics, University of Strathclyde, U.K. He has been with the Faculty of Science and Institute for Information and System Sciences, Xi'an Jiaotong University, since 1982, where he became Associate Professor in 1987 and Full Professor in 1991, and now serves as Director of the Institute for Information and System Sciences. His current research interests include nonlinear functional analysis, mathematical foundation of information technology, and computational intelligence.

Dr. Xu was given the second prize of National Natural Science Award of China in 2007. He now serves as Chief Scientist of one National Basic Research Project of China (973 Project).



Rui Zhang received the B.S. and M.S. degrees in mathematics from Northwest University, Xi'an, China, in 1994 and 1997, respectively. She is currently working towards the Ph.D. degree at the Institute for Information and System Sciences, Xi'an Jiaotong University, Xi'an, China.

She has been with the Department of Mathematics, Northwest University, since 1997, where she was promoted to an Associate Professor in 2007. From August 2004 to January 2005, she was a Visiting Scholar at the Department of Mathematics, University of Illinois at Champaign-Urbana, Urbana. Her current research interests include optimization theory and application, neural networks, and evolutionary computations.

Wen-Feng Jing received the B.S. and M.S. degrees in mathematics from Xi'an Jiaotong University, Xi'an, China, in 1988 and 2002, respectively, where he is currently working towards the Ph.D. degree at the Institute for Information and System Sciences.

He has been with the School of Science, Xi'an Jiaotong University, where he was promoted to an Associate Professor in 2000. His current research interests include data mining and machine learning.