DE GRUYTER
OPEN

JACSM 2016, Vol. 8, No. 1, pp. 5 - 15

10.1515/jacsm-2016-0001

CONVERGENCE ANALYSIS OF AN IMPROVED EXTREME LEARNING MACHINE BASED ON GRADIENT DESCENT METHOD

Yusong Liu[1,2], Zhixun Su[1], Bingjie Zhang[2], Xiaoling Gong[3], Zhaoyang Sang[2]

[1] School of Mathematical Sciences,
Dalian University of Technology, Dalian 116024, China
yangguolingfwz@163.com
ysliu1758@163.com

[2] College of Science,
China University of Petroleum (Huadong), Qingdao 266580, China

[3] College of Information and Control Engineering,
China University of Petroleum (Huadong), Qingdao 266580, China

Abstract

Extreme learning machine (ELM) is an efficient algorithm, but it requires more hidden nodes than the BP algorithms to reach the matched performance. Recently, an efficient learning algorithm, the upper-layer-solution-unaware algorithm (USUA), is proposed for the single-hidden layer feed-forward neural network. It needs less number of hidden nodes and testing time than ELM. In this paper, we mainly give the theoretical analysis for USUA. Theoretical results show that the error function monotonously decreases in the training procedure, the gradient of the error function with respect to weights tends to zero (the weak convergence), and the weight sequence goes to a fixed point (the strong convergence) when the iterations approach positive infinity. An illustrated simulation has been implemented on the MNIST database of handwritten digits which effectively verifies the theoretical results..

Key words: Neural networks, Monotonicity, Weak convergence, Strong convergence, USUA, MNIST.

1 Introduction

Neural network has been a hot topic recently in many fields, such as cognitive science, prediction, classification, computational intelligence. The back-propagation (BP) algorithm is one of the most widely used techniques for training feed-forward neural networks (FNN), which was separately proposed by Werbos [1] and Rumelhart et al.[2]. The BP algorithm attempts to minimize the least squared error of objective function, which is defined by the differences between the actual outputs and the desired outputs [3]. In BP algorithm, all the weights of FNN need to be tuned along the negative gradient direction of the error function using the gradient descent method.

The BP algorithm for FNN has the ability of approximating nonlinear functions directly from the input samples. However, the training procedure of the BP algorithm is usually very time consuming. The reasons come from two aspects: (1) the gradient-based learning algorithms are used in training the neural networks, and (2) all weights of the neural networks are tuned in each iteration.

To overcome these shortcomings, Huang et al.[4] proposed a novel learning algorithm called extreme learning machine (ELM) for single-hidden layer feed-forward neural networks (SHLFN), which randomly chooses hidden weights and determines the output weights of SHLFN.

Specifically speaking, in ELM algorithm, the weights connecting the input and hidden layers are selected randomly, and the weights connecting the hidden and output layers are only calculated using the pseudo inverse once. There is no iteration step in the training procedure. In addition, the training speed of ELM is much faster than that of the BP learning algorithms when reaching the comparable performance.

Although ELM can be trained efficiently, it requires more hidden nodes than the BP algorithms for the trained neural networks. This apparently increases testing time which does not effectively work well in real applications.

Yu et al.[5] proposed a series of efficient learning algorithms for SHLFN. The main idea is that, giving the initial weights of FNN, the weights connecting the input and hidden layers are tuned in the negative gradient direction along which the square error is reduced the most, and then the weights connecting the hidden and output layers are calculated using the pseudo inverse. Numerical experiment shows that the proposed algorithms in [5] need less number of hidden nodes and testing time than ELM.

Unfortunately, there is little theoretical analysis to guarantee the convergent behavior during training. In this paper, we rigorously prove the theoretical results for the upper-layer-solution-unaware algorithm (USUA) proposed by Yu et al.[5]. The error function monotonously decreases during training. The weak convergence and the strong convergence show that the gradient of the error function goes to zero, and the weight sequence goes to a unique fixed point, respectively. Numerical experiment on the MNIST database of handwritten digits [6] verifies these theoretical results.

The rest of this paper is organized as follows. Section 2 gives a brief introduction to USUA. Section 3 presents the main theoretical results of USUA. Section 4 rigorously proves these theoretical results. A numerical experiment is simulated in Section 5.

2 USUA

The SHLFN is considered. The number of nodes of the input, hidden and output layers are set to be *D*, *L* and *C*, respectively.

The matrix $\mathbf{W}=(\mathbf{w}_1,\mathbf{w}_2,\cdots,\mathbf{w}_L)\in R^{D\times L}$ represents the weight connections between the input and hidden layers, where $\mathbf{w}_i=(w_{1i},w_{2i},\cdots,w_{Di})^T\in R^D$ is the weight vector connecting the input nodes and the i-th hidden node. $\mathbf{U}=(\mathbf{u}_1,\mathbf{u}_2,\cdots,\mathbf{u}_C)\in R^{L\times C}$ denotes the weight matrix connecting the hidden and output layers, where $\mathbf{u}_i=(u_{1i},u_{2i},\cdots,u_{Li})^T\in R^L$ is the weight vector connecting the hidden nodes and the i-th output node. For simplicity, $\mathbf{W}$ is rewritten as $\mathbf{V}=(\mathbf{w}_1{}^T,\mathbf{w}_2{}^T,\cdots,\mathbf{w}_L{}^T)^T\in R^{DL}$.

Let $g,f:R\to R$ be the given activation functions of the hidden and output layers, respectively. For any given vector $\mathbf{z}=(z_1,z_2,\cdots,z_L)^T\in R^L$, the vector valued function is introduced, denoting as

$$\mathbf{G}(\mathbf{z})=(g(z_1),g(z_2),\cdots,g(z_L))^T\in R^L. \quad (1)$$

For any given output vector $\mathbf{x}\in R^D$, the actual output vector of the neural network is $\mathbf{y}\in R^C$, i.e.

$$\mathbf{y}=f(\mathbf{U}^T\mathbf{G}(\mathbf{W}^T\mathbf{x})).$$

Yu et al.[5] present an efficient and effective algorithm for training SHLFN named USUA. The basic idea of USUA is as follows: when the initial value of **V** and **U** are given, the weight matrix **U** is then fixed, and the weight matrix **V** is updated by using the gradient descend method until it reaches the stop criteria. Then, **U** is calculated using the pseudo inverse. The detailed description is as follows.

Given a training sample set with ***N*** samples, $\mathbf{X}=\{\mathbf{x}_i\}_1^N$ is the set of the input vectors, $\mathbf{T}=\{\mathbf{t}_i\}_1^N$ is the set of the corresponding ideal outputs, and the actual output of the output layer are $\mathbf{Y}=\{\mathbf{y}_i\}_1^N$, where $\mathbf{x}_i\in R^D$, $\mathbf{t}_i\in R^C$, $\mathbf{y}_i\in R^C$. The objective function of the neural networks is defined as follows,

$$\begin{aligned}E(\mathbf{V})&=\frac{1}{2}\|\mathbf{Y}-\mathbf{T}\|_{\mathrm{F}}^2=\frac{1}{2}\sum_{i=1}^{N}\|\mathbf{y}_i-\mathbf{t}_i\|^2\\&=\frac{1}{2}\sum_{i=1}^{N}\left\|f(\mathbf{U}^T\mathbf{G}(\mathbf{W}^T\mathbf{x}_i))-\mathbf{t}_i\right\|^2\end{aligned}$$

$$=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{C}(f(\mathbf{u}_j^T\mathbf{G}(\mathbf{W}^T\mathbf{x}_i))-t_{ji})^2$$

$$=\sum_{i=1}^{N}\sum_{j=1}^{C}f_{ji}(\mathbf{u}_j^T\mathbf{G}(\mathbf{W}^T\mathbf{x}_i)), \tag{2}$$

where $\|\cdot\|_{\mathrm{F}}$ and $\|\cdot\|$ stand for the Frobenius norm of matrix and the Euclidean norm of vector, respectively, and $f_{ji}(s)=\frac{1}{2}(f(s)-t_{ji})^2, s\in R$.

The gradients of the error function $E(\mathbf{V})$ with respect to $\mathbf{w}_k$ $(k=1,2,\cdots,L)$ are

$$E_{\mathbf{w}_k}(\mathbf{V})=\sum_{i=1}^{N}\sum_{j=1}^{C}f_{ji}^{'}(\mathbf{u}_j^T\mathbf{G}(\mathbf{W}^T\mathbf{x}_i))u_{kj}g^{'}(\mathbf{w}_k^T\mathbf{x}_i)\mathbf{x}_i\,. \tag{3}$$

Denote

$$E_{\mathbf{V}}(\mathbf{V})=((E_{\mathbf{w}_1}(\mathbf{V}))^T,(E_{\mathbf{w}_2}(\mathbf{V}))^T,\cdots,(E_{\mathbf{w}_L}(\mathbf{V}))^T)^T\,. \tag{4}$$

For any given initial weight vector $\mathbf{V}^0$ and $\mathbf{U}$, $\mathbf{V}$ can be iterated by the following formula

$$\mathbf{V}^{n+1}=\mathbf{V}^n+\Delta\mathbf{V}^n, \quad n=0,1,2,\cdots, \tag{5}$$

where $\Delta\mathbf{V}^n$=$((\Delta\mathbf{w}_1^n)^T,(\Delta\mathbf{w}_2^n)^T,\cdots,(\Delta\mathbf{w}_L^n)^T)^T$, and

$$\Delta\mathbf{w}_k^n=-\eta\sum_{i=1}^{N}\sum_{j=1}^{C}f_{ji}^{'}(\mathbf{u}_j^T\mathbf{G}(\mathbf{W}^T\mathbf{x}_i))u_{kj}g^{'}(\mathbf{w}_k^T\mathbf{x}_i)\mathbf{x}_i\,, \tag{6}$$

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

At last, **U** is calculated using the pseudo inverse.

3 The main convergence results

To analyze the convergence of USUA, the following assumptions are needed.

(A1)The activation functions g and f satisfy that, $|g(s)|$, $|f(s)|$, $|g^{'}(s)|$, $|f^{'}(s)|$, $|g^{''}(s)|$ and $|f^{''}(s)|$ are all uniformly bounded for any $s\in R$.

(A2) There are finitely many points in the set $\Omega_0=\{\mathbf{V}\in\Omega:E_{\mathbf{V}}(\mathbf{V})=0\}$, where Ω is a bounded closed region.

Theorem 1. Assume that assumption (A1) is valid, and the learning rate η satisfies the formula (21) behind. Then, for any arbitrary initial weight vector $\mathbf{V}^0$, the sequence $\{E(\mathbf{V}^n)\}$ monotonously decreases, i.e.

$$E(\mathbf{V}^{n+1}) \le E(\mathbf{V}^n)\,; \tag{7}$$

there exists $E^* \ge 0$, such that

$$\lim_{n\to\infty} E(\mathbf{V}^n) = E^*\,; \tag{8}$$

and the weak convergence result holds,

$$\lim_{n\to\infty} \left\| E_{\mathbf{V}}(\mathbf{V}^n) \right\| = 0\,. \tag{9}$$

In addition, if assumption (A2) is also valid, then the strong convergence result holds, i.e. there exists $\mathbf{V}^* \in \Omega_0$, such that

$$\lim_{n\to\infty} \mathbf{V}^n = \mathbf{V}^*. \tag{10}$$

4 The Proofs

The proofs of the convergence results (Theorem 1) are presented as follows. Firstly, two useful lemmas are given. For sake of consistency, denote

$$\Delta\mathbf{w}_k^n = \mathbf{w}_k^{n+1} - \mathbf{w}_k^n\,, \tag{11}$$

$$\mathbf{G}^{n,i} = \mathbf{G}((\mathbf{W}^n)^T \mathbf{x}_i),\ \boldsymbol{\varphi}^{n,i} = \mathbf{G}^{n+1,i} - \mathbf{G}^{n,i}\,. \tag{12}$$

Lemma 1. If assumption (A1) is valid, then there exist $c_1 > 0$ and $c_2 > 0$, satisfying

$$\left\|\boldsymbol{\varphi}^{n,i}\right\|^2 \le c_1 \sum_{k=1}^{L} \left\|\Delta\mathbf{w}_k^n\right\|^2, \quad i = 1,2,\cdots,N,\ \ n = 1,2,\cdots, \tag{13}$$

$$\left|f_{ji}^{'}(s)\right| \le c_2, \left|f_{ji}^{''}(s)\right| \le c_2,\ s \in R,\ i = 1,2,\cdots,N,\ j = 1,2,\cdots,C\,. \tag{14}$$

Proof. According to assumption (A1) and the Taylor expansion, we get

$$\left\|\boldsymbol{\varphi}^{n,i}\right\|^2 = \left\|\mathbf{G}^{n+1,i} - \mathbf{G}^{n,i}\right\|^2 = \left\|\begin{pmatrix} g((\mathbf{w}_1^{n+1})^T \mathbf{x}_i) - g((\mathbf{w}_1^n)^T \mathbf{x}_i) \\ g((\mathbf{w}_2^{n+1})^T \mathbf{x}_i) - g((\mathbf{w}_2^n)^T \mathbf{x}_i) \\ \vdots \\ g((\mathbf{w}_L^{n+1})^T \mathbf{x}_i) - g((\mathbf{w}_L^n)^T \mathbf{x}_i) \end{pmatrix}\right\|^2$$

$$= \left\|\begin{pmatrix} g^{'}(s_{1,i,n})\Delta(\mathbf{w}_1^n)^T \mathbf{x}_i \\ g^{'}(s_{2,i,n})\Delta(\mathbf{w}_2^n)^T \mathbf{x}_i \\ \vdots \\ g^{'}(s_{L,i,n})\Delta(\mathbf{w}_L^n)^T \mathbf{x}_i \end{pmatrix}\right\|^2$$

$$\leq (\sup_{s\in R}\left|g^{'}(s)\right|\max_{1\leq i\leq N}\left\|\mathbf{x}_i\right\|)^2\sum_{k=1}^{L}\left\|\Delta\mathbf{w}_k^n\right\|^2$$

$$= c_1\sum_{k=1}^{L}\left\|\Delta\mathbf{w}_k^n\right\|^2 ,$$

where $c_1 = (\sup_{s\in R}\left|g^{'}(s)\right|\max_{1\leq i\leq N}\left\|\mathbf{x}_i\right\|)^2$, and $s_{k,i,n}(k=1,2,\cdots,L)$ lies between $(\mathbf{w}_k^{n+1})^T\mathbf{x}_i$ and $(\mathbf{w}_k^n)^T\mathbf{x}_i$.

By the expression of $f_{ji}(s)$ and assumption (A1), it is easily known that

$$\left|f_{ji}^{'}(s)\right|\leq c_2, \left|f_{ji}^{''}(s)\right|\leq c_2,\ i=1,2,\cdots,N,\ \ j=1,2,\cdots,C,\ \ s\in R,$$

where $c_2 = \max\{\sup_{s\in R}\left|(f(s)-t_{ji})f^{'}(s)\right|, \sup_{s\in R}\left|(f^{'}(s))^2+(f(s)-t_{ji})f^{''}(s)\right|\}$.

The following lemma is the same as Theorem 14.1.5 [7], therefore we only list it below without proof.

Lemma 2. [7] Let $F:\Omega\subset R^n\to R^m(n,m\geq 1)$ be continuous on a bounded closed region $\Omega\subset R^n$, and $\Omega_0=\{\mathbf{z}\in\Omega:F(\mathbf{z})=0\}$ be a finite set. Let $\{\mathbf{z}^k\}\subset\Omega$ be a sequence satisfying

(1) $\lim_{k\to\infty}F(\mathbf{z}^k)=0$,

(2) $\lim_{k\to\infty}\left\|\mathbf{z}^{k+1}-\mathbf{z}^k\right\|=0$,

then, there exists a $\mathbf{z}^*\in\Omega_0$ such that $\lim_{k\to\infty}\mathbf{z}^k=\mathbf{z}^*$.

Next, the proofs for (7)-(10) are successively presented as follows.

Proof for (7).

By (2) and the Taylor expansion, we have

$$\begin{aligned}
&E(\mathbf{V}^{n+1})-E(\mathbf{V}^n)\\
&=\sum_{i=1}^{N}\sum_{j=1}^{C}[f_{ji}(\mathbf{u}_j^T\mathbf{G}^{n+1,i})-f_{ji}(\mathbf{u}_j^T\mathbf{G}^{n,i})]\\
&=\sum_{i=1}^{N}\sum_{j=1}^{C}[f_{ji}^{'}(\mathbf{u}_j^T\mathbf{G}^{n,i})\mathbf{u}_j^T(\mathbf{G}^{n+1,i}-\mathbf{G}^{n,i})+\frac{1}{2}f_{ji}^{''}(\tilde{s}_{n,i})(\mathbf{u}_j^T(\mathbf{G}^{n+1,i}-\mathbf{G}^{n,i}))^2]\\
&=\delta_0+\delta_1, \qquad (15)
\end{aligned}$$

where $\delta_0 = \sum_{i=1}^{N}\sum_{j=1}^{C} f_{ji}^{'}(\mathbf{u}_j^T\mathbf{G}^{n,i})\mathbf{u}_j^T(\mathbf{G}^{n+1,i} - \mathbf{G}^{n,i})$,

$\delta_1 = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{C} f_{ji}^{''}(\tilde{s}_{n,i})(\mathbf{u}_j^T(\mathbf{G}^{n+1,i} - \mathbf{G}^{n,i}))^2$, and $\tilde{s}_{n,i}$ lies between $\mathbf{u}_j^T\mathbf{G}^{n+1,i}$ and $\mathbf{u}_j^T\mathbf{G}^{n,i}$.

By the Taylor expansion and (6),

$$\delta_0 = \sum_{i=1}^{N}\sum_{j=1}^{C} f_{ji}^{'}(\mathbf{u}_j^T\mathbf{G}^{n,i})[\sum_{k=1}^{L} u_{kj}(g((\mathbf{w}_k^{n+1})^T\mathbf{x}_i) - g((\mathbf{w}_k^{n})^T\mathbf{x}_i))]$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{C} f_{ji}^{'}(\mathbf{u}_j^T\mathbf{G}^{n,i})[\sum_{k=1}^{L} u_{kj}g^{'}((\mathbf{w}_k^{n})^T\mathbf{x}_i)(\Delta\mathbf{w}_k^n)^T\mathbf{x}_i + \frac{1}{2}\sum_{k=1}^{L} u_{kj}g^{''}(\bar{s}_{k,i,n})((\Delta\mathbf{w}_k^n)^T\mathbf{x}_i)^2]$$

$$= -\frac{1}{\eta}\sum_{k=1}^{L}\left\|\Delta\mathbf{w}_k^n\right\|^2 + \delta_2, \tag{16}$$

where $\delta_2 = \frac{1}{2}\sum_{k=1}^{L}\sum_{i=1}^{N}\sum_{j=1}^{C} f_{ji}^{'}(\mathbf{u}_j^T\mathbf{G}^{n,i})u_{kj}g^{''}(\bar{s}_{k,i,n})((\Delta\mathbf{w}_k^n)^T\mathbf{x}_i)^2$, and $\bar{s}_{k,i,n}$ lies between $(\mathbf{w}_k^{n+1})^T\mathbf{x}_i$ and $(\mathbf{w}_k^{n})^T\mathbf{x}_i$.

By (15) and (16),

$$E(\mathbf{V}^{n+1}) - E(\mathbf{V}^{n}) = -\frac{1}{\eta}\sum_{k=1}^{L}\left\|\Delta\mathbf{w}_k^n\right\|^2 + \delta_2 + \delta_1. \tag{17}$$

As (15), $\mathbf{U}$ is fixed and the triangle inequality,

$$\delta_1 \le \frac{1}{2}c_2\sum_{i=1}^{N}\sum_{j=1}^{C}\left\|\mathbf{u}_j^T\boldsymbol{\varphi}^{n,i}\right\|^2$$

$$\le \frac{1}{2}c_2\sum_{i=1}^{N}\sum_{j=1}^{C}(\max_{1\le j\le C}\left\|\mathbf{u}_j\right\|)^2 c_1\sum_{k=1}^{L}\left\|\Delta\mathbf{w}_k^n\right\|^2$$

$$= \frac{1}{2}c_2 NCc_1(\max_{1\le j\le C}\left\|\mathbf{u}_j\right\|)^2\sum_{k=1}^{L}\left\|\Delta\mathbf{w}_k^n\right\|^2$$

$$= c_3\sum_{k=1}^{L}\left\|\Delta\mathbf{w}_k^n\right\|^2, \tag{18}$$

where $c_3 = \frac{1}{2}c_2 NCc_1(\max_{1\le j\le C}\left\|\mathbf{u}_j\right\|)^2$.

By assumption (A1) and (14),

$$\delta_2 \le \frac{1}{2}\sum_{k=1}^{L}\sum_{i=1}^{N}\sum_{j=1}^{C} c_2\max_{1\le j\le C}\left\|\mathbf{u}_j\right\|\sup_{s\in R}\left|g^{''}(s)\right|\max_{1\le i\le N}\left\|\mathbf{x}_i\right\|^2\left\|\Delta\mathbf{w}_k^n\right\|^2$$

$$=\frac{1}{2}NCc_2\max_{1\le j\le C}\left\|\mathbf{u}_j\right\|\sup_{s\in R}\left|g^{"}(s)\right|\max_{1\le i\le N}\left\|\mathbf{x}_i\right\|^2\sum_{k=1}^{L}\left\|\Delta\mathbf{w}_k^n\right\|^2$$

$$=c_4\sum_{k=1}^{L}\left\|\Delta\mathbf{w}_k^n\right\|^2, \qquad (19)$$

where $c_4=\frac{1}{2}NCc_2\max_{1\le j\le C}\left\|\mathbf{u}_j\right\|\sup_{s\in R}\left|g^{"}(s)\right|\max_{1\le i\le N}\left\|\mathbf{x}_i\right\|^2$.

Therefore, by (17), (18) and (19)

$$\begin{aligned}&E(\mathbf{V}^{n+1})-E(\mathbf{V}^n)\\ &\le-\frac{1}{\eta}\sum_{k=1}^{L}\left\|\Delta\mathbf{w}_k^n\right\|^2+c_4\sum_{k=1}^{L}\left\|\Delta\mathbf{w}_k^n\right\|^2+c_3\sum_{k=1}^{L}\left\|\Delta\mathbf{w}_k^n\right\|^2\\ &=-(\frac{1}{\eta}-c_5)\sum_{k=1}^{L}\left\|\Delta\mathbf{w}_k^n\right\|^2\\ &=-\alpha\sum_{k=1}^{L}\left\|\Delta\mathbf{w}_k^n\right\|^2,\end{aligned} \qquad (20)$$

where $c_5=c_3+c_4$, $\alpha=\frac{1}{\eta}-c_5$. Set

$$0<\eta<\frac{1}{c_5}, \qquad (21)$$

then, $E(\mathbf{V}^{n+1})\le E(\mathbf{V}^n)$. The monotonicity is proved.

Proof for (8).

For any $n=0,1,2,\cdots$, $E(\mathbf{V}^n)\ge 0$. Then, by (7), the sequence $\{E(\mathbf{V}^n)\}$ monotonously decreases. Therefore, there exists $E^*\ge 0$ such that $\lim_{n\to\infty}E(\mathbf{V}^n)=E^*$.

Proof for (9).

By (20), we obtain

$$\begin{aligned}E(\mathbf{V}^{n+1})&\le E(\mathbf{V}^n)-\alpha\sum_{k=1}^{L}\left\|\Delta\mathbf{w}_k^n\right\|^2\\ &\le E(\mathbf{V}^{n-1})-\alpha\sum_{k=1}^{L}\left\|\Delta\mathbf{w}_k^{n-1}\right\|^2-\alpha\sum_{k=1}^{L}\left\|\Delta\mathbf{w}_k^n\right\|^2\\ &\le\cdots\le E(\mathbf{V}^0)-\alpha\sum_{i=0}^{n}\sum_{k=1}^{L}\left\|\Delta\mathbf{w}_k^i\right\|^2.\end{aligned}$$

For any $n\ge 0$, we have $E(\mathbf{V}^n)\ge 0$. Therefore, $\alpha\sum_{i=0}^{n}\sum_{k=1}^{L}\left\|\Delta\mathbf{w}_k^i\right\|^2\le E(\mathbf{V}^0)$.

By (3), (4) and (6), taking $n \to \infty$, and changing indexes,

$$\alpha \sum_{n=0}^{\infty} \sum_{k=1}^{L} \left\| \Delta \mathbf{w}_k^n \right\|^2 = \alpha \eta^2 \sum_{n=0}^{\infty} \left\| E_{\mathrm{V}}(\mathbf{V}^n) \right\|^2 \leq E(\mathbf{V}^0) < \infty .$$

Then, we have $\lim_{n\to\infty} \left\| E_{\mathrm{V}}(\mathbf{V}^n) \right\| = 0$. The weak convergence is proved.

Proof for (10).

By (3)-(6),

$$\left\| \mathbf{V}^{n+1} - \mathbf{V}^n \right\| = \left\| \Delta \mathbf{V}^n \right\| = \eta \left\| E_{\mathrm{V}}(\mathbf{V}^n) \right\| .$$

Thus, using (9), we get

$$\lim_{n\to\infty} \left\| \mathbf{V}^{n+1} - \mathbf{V}^n \right\| = 0 .$$

By (A2), the conditions of lemma 2 are valid. Therefore, there exists $\mathbf{V}^* \in \Omega_0$ satisfying $\lim_{n\to\infty} \mathbf{V}^n = \mathbf{V}^*$.The strong convergence is proved.

5 Numerical experiment

The MNIST database of handwritten digits contains 60,000 training samples and 10,000 testing samples. Each digital image has been normalized to an image 28×28 pixels, and expanded as a 784×1 vector. The elements of these digital vectors are the integer numbers between 0-255.

According to the property of MNIST, we construct a network model whose structure is set to be 784-128-10. The learning rate is selected as a constant 0.0007. The activation functions of hidden and output layers are with the common sigmoid function $g(x) = \frac{1}{1+e^{-x}}$ and the linear function, respectively. The initial weights are randomly assigned in the interval $[-1,1]$. The stop criteria are set to be: 1,000 training epochs or the error below 0.01.

Figure 1 and Figure 2 display the classification ability of the USUA on training and testing samples. To show the details clearly, the accuracies are recorded for each training epoch. We observe that the USUA has the similar performance on both training and testing samples. In addition, the two curves drastically increase in the early training stage and then maintain with a stable status.

In Figure 3, it shows the error values of each training epoch. Corresponding to the training performance in Figure 2, the errors sharply decrease in the early training epochs, and the approach the minimum. This effectively verifies the monotonicity of error function which is proved in Theorem 1.

For the last Figure 4, the norms of the gradient of error function with respect to weight vectors have been graphed along with epochs. Although the curve shows the oscillation behavior in the training process, it still demonstrates that the norms tend to small values near zero along with the increasing epochs. This then illustrates the proved weak convergence of USUA in Theorem 1.

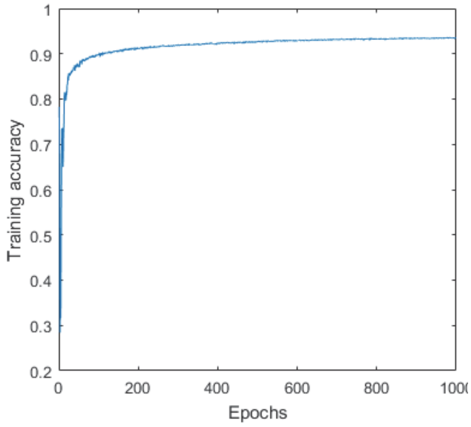

Figure 1. The curve of training accuracy

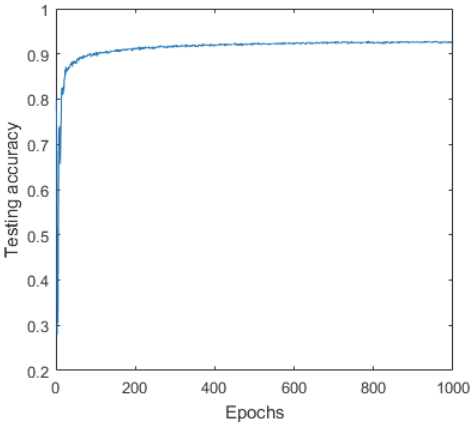

Figure 2. The curve of testing accuracy

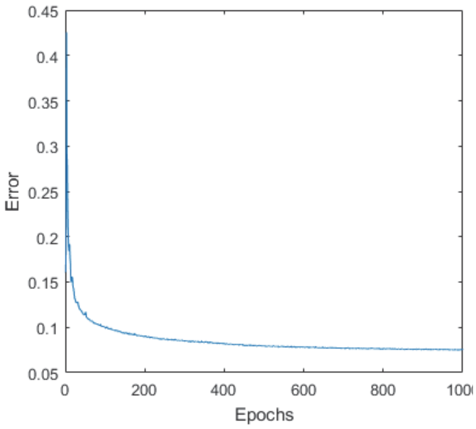

Figure 3. The curve of error function

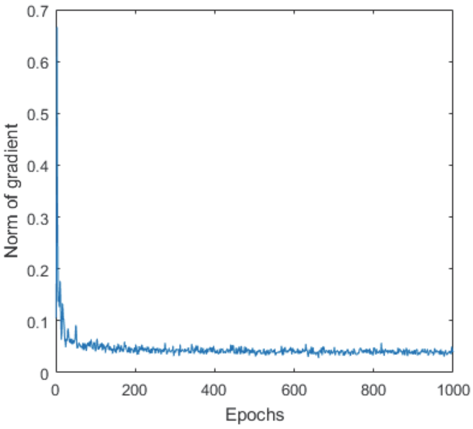

Figure 4. The norms of the gradient of error function with respect to weight vectors

6 Conclusion

In this paper, we mainly rigorously prove the theoretical results of USUA proposed by Yu et al.[5], including the monotonicity of error function, the weak and strong convergence. The error function monotonously decreases in the training procedure. The weak and strong convergence indicate that the gradient of the error function with respect to weights tends to zero and the weight sequence goes to a fixed point when the iterations approach positive infinity, respectively. Numerical experiment on the MNIST database of handwritten digits support these theoretical results.

Acknowledgments

The authors wish to thank the anonymous reviewers for careful error proofing of the manuscript and many insightful comments and suggestions which greatly improved this work.

This project was supported in part by the National Natural Science Foundation of China (No. 61305075, 61173103, 61572099, 61320106008, 91230103), the China Postdoctoral Science Foundation (No. 2012M520624), Natural Science Foundation of Shandong Province (No. ZR2013FQ004, ZR2013DM015), the Specialized Research Fund for the Doctoral Program of Higher Education of China (No. 20130133120014) and the Fundamental Research Funds for the Central Universities (No. 15CX05053A, 15CX08011A).

References

1. Werbos, P. J., 1974, *Beyond regression: new tools for prediction and analysis in the behavioral sciences*, Ph.D. thesis, Harvard University, Cambridge, MA
2. Rumelhart D. E., Hinton G. E., Williams R. J., 1986, *Learning representations by back-propagating errors*, Nature, Vol. 323, pp. 533-536
3. J.H. Goodband, O.C.L. Haas, J.A. Mills, 2008, *A comparison of neural network approaches for on-line prediction in IGRT*, Medical Physics, Vol. 35, No. 3, pp. 1113–1122
4. Huang, G.-B., Zhu, Q.-Y., Siew, C.-K., 2006, *Extreme learning machine: theory and applications*, Neurocomputing, Vol. 70, No. 1-3, pp. 489-501
5. D. Yu and L. Deng, 2012, *Efficient and effective algorithms for training single-hidden-layer neural networks*, Pattern Recognition Letters, Vol. 33, No. 5, pp. 554–558
6. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, 1998, *Gradient-based learning applied to document recognition*, Proceedings of the IEEE, Vol. 86, No. 11, pp. 2278-2324
7. Y. Yuan, W. Sun, 2001, *Optimization Theory and Methods*, Science Press, Beijing