International Journal of Innovation and Applied Studies
ISSN 2028-9324 Vol. 9 No. 3 Nov. 2014, pp. 1056-1063

http://www.ijias.issr-journals.org/

Convergence of Offline Gradient Method with Smoothing $L_{1/2}$ Regularization for Two-layer of Neural Network

***Khidir Shaib Mohamed*[1-2] *and Yousif Shoaib Mohammed*[3-4-5]**

[1]School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, PR China

[2]Mathematical Department, College of Science, Dalanj University, Dalanj, Sudan

[3]Department of Physics, College of Science & Art, Qassim University, Oklat Al- Skoor, P.O.Box: 111, Saudi Arabia

[4]Physics Department, College of Education, Dalanj University, Dalanj, Sudan

[5]Physics Department, Africa City for Technology – Khartoum, Sudan

ABSTRACT: In this paper, we study the convergence of offline gradient method with smoothing $L_{1/2}$ regularization penalty for training multi-output feed forward neural networks. The monotonicity of the error function and weight boundedness for the offline gradient with smoothing $L_{1/2}$ regularization. the usual $L_{1/2}$ regularization term involves absolute value and is not differentiable at the origin. The key point of this paper is modify the usual $L_{1/2}$ regularization term by smoothing it at the origin are presented, the convergence results are proved, which will be very meaningful for theoretical research or applications on multi – output neural networks.

KEYWORDS: feed forward neural network; offline gradient method; smoothing $L_{1/2}$ regularization; boundedness; convergence.

1 INTRODUCTION

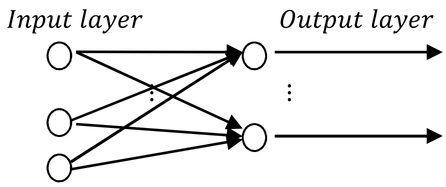
Feed forward neural networks (FNN) have been widely used in much application [1 – 9]. The Penalty (regularization) term is often introduced into the network training algorithms so as to control the magnitude of the weights and to improve the generalization performance of the network [10 - 13]. A commonly used penalty term added to the standard error function is a term proportional to the norm of the weights. The effectiveness of this penalty has been tested on many problems such as the monks problem, etc. An especially $L_{1/2}$ regularization term is introduced into the batch gradient learning algorithm for pruning of FNN, the usual $L_{1/2}$ regularization term is not smooth at the origin, which causes difficulty in the convergence analysis and, more importantly. Oscillation in the numerical computation as observed in the numerical experiments [14]. In [15], some convergence results are given for feed forward neural networks with $L_{1/2}$ regularization penalty, where the learning fashion of training examples is gradient algorithm learning. The key for the convergence results of the error function is decreasing monotonically, and the online gradient method with $L_{1/2}$ smoothing regularization term is deterministically convergent. As a simple example, the convergence for two- layer feed forward neural network is discussed in [16]. The convergence of the online and batch gradient algorithm with a penalty term for feed forward neural network has been also discussed [17 - 19]. These results are of global nature that they are valid for any arbitrarily given initial values of the weights. In addition, multi-output feed forward neural network is widely used in classification problems, and the convergence of multi-output neural network is very meaningful.

In this paper, we study a multi-output BP neural network with $L_{1/2}$ regularization penalty term and define a relation formula between the penalty parameter and the learning rate parameter, then use it to prove the weak and strong convergences of the offline gradient algorithm with $L_{1/2}$ regularization penalty. We note that usual $L_{1/2}$ regularization penalty term is not smooth at the origin, so difficulty to proof main results, for which we suppose that $L_{1/2}$ regularization smoothing in the origin error function, then it come easy to prove, our main results. Additionally, the boundness of the new error function with $L_{1/2}$ regularization penalty is also guaranteed.

The rest of this paper is arranged as follows. The offline gradient method with smoothing $L_{1/2}$ regularization penalty term is described in section 2. In section 3, the convergence results of offline gradient method with smoothing $L_{1/2}$ regularization are presented. In section 4, we display a brief summary of our present work.

2 ALGORITHM DESCRIPTION

In this section, we introduction a two- layer network consisting of p input layers, n output layers. Fig. 1 illustrates the structure of two-layer multi-output feed-forward neural network



$\boldsymbol{Fig.1\ structure\ of\ two-layer\ Multi-output\ feedforward\ neural\ network}$

Denote the $\omega = (\omega_{ab})_{IP}$ and $\omega_i = (\omega_{i1}, \omega_{i2}, \dots, \omega_{ip})$ $(1 \le i \le n)$ and the transfer function by $g: \mathbb{R} \to \mathbb{R}$, is a sigmoid function. Let $\{x^j, o^j\}_{j=1}^J \subset \mathbb{R}^p \times \mathbb{R}$ is a given of training samples.

For each input $\xi \in \mathbb{R}^p$, then the actual output is computed by

$$\xi = g(\omega \cdot \xi) \tag{1}$$

2.1 OFFLINE GRADIENT METHOD WITH $L_{1/2}$ REGULARIZATION

In supervised training of neural networks, synaptic weights are usually updated by an iterative algorithm which searches for the minimum of a cost function. A popular choice of the cost function is

$$\tilde{E}(\omega) = \frac{1}{2}\sum_{j=1}^{J}(o^j - \xi^j) = \frac{1}{2}\sum_{j=1}^{J}\sum_{i=1}^{n}\left(o^j - g(\omega_i \cdot \xi^j)\right)^2 \tag{2}$$

By adding a $L_{1/2}$ regularization penalty term, the modified cost function takes the form (cf. [6, 12])

$$\begin{aligned} E(\omega) &= \frac{1}{2}\sum_{j=1}^{J}\sum_{i=1}^{n}\left(o^j - g_{ji}(\omega_i \cdot \xi^j)\right)^2 + \lambda\sum_{i=1}^{n}\sum_{k=1}^{p}|\omega_{ik}|^{1/2} \\ &= \sum_{j=1}^{J}\sum_{i=1}^{n} g_{ji}(\omega_i \cdot \xi^j) + \lambda\sum_{i=1}^{n}\sum_{k=1}^{p}|\omega_{ik}|^{1/2} \end{aligned} \tag{3}$$

Where $g_j(t) \coloneqq \frac{1}{2}(o^j - g(t))^2$ and $\lambda > 0$ is a penalty parameter. Then the gradient function with respect to ω_{ik}($i = 1,2,\dots,n;$ $k = 1,2,\dots,p$) is

$$E_{\omega_{ik}}(\omega) = \sum_{j=1}^{J} g'_{ji}(\omega_i \cdot \xi^j)\xi_k^j + \frac{\lambda\, sgm\,(\omega_{ik})}{2\,|\omega_{ik}|^{1/2}} \tag{4}$$

Let ω^0 be arbitrary initial weight. The offline gradient method with $L_{1/2}$ regularization term updates the weights $\{\omega^m\}$ iteratively by

$$\omega_{ik}^{mJ+j} = \omega_{ik}^{mJ+j-1} - \Delta_{\mathrm{j}}^{\mathrm{m}} \omega_{ik}^{mJ+j-1}, \qquad m = 0,1,2, \ldots. \tag{5}$$

Where

$$\Delta_{\mathrm{j}}^{\mathrm{m}} \omega_{ik}^{mJ+j-1} = -\eta_m \nabla_{\omega_{ik}} E\left(\omega_{ik}^{mJ+j-1}\right) = -\eta_m E_{\omega_{ik}}\left(\omega_{ik}^{mJ+j-1}\right)$$

$$= -\eta_m \left(\sum_{j=1}^{J} g'_{ji}\left(\omega_i^{mJ+j-1} \cdot \xi^j\right) \xi_k^j + \frac{\lambda \, sgm\left(\omega_{ik}^{mJ+j-1}\right)}{2\left|\omega_{ik}^{mJ+j-1}\right|^{1/2}} \right) \tag{6}$$

Here, $\nabla_{\omega_{ik}} E(\omega^{mJ+j-1})$ is the gradient of $E(\omega^{mJ+j-1})$ with respect to ω_{ik} $(i = 1,2, \ldots, n;\ k = 1,2, \ldots, p)$ and $\eta_m > 0$ is the learning rate.

2.2 OFFLINE GRADIENT METHOD WITH SMOOTHING $L_{1/2}$ REGULARIZATION

A modified $L_{1\backslash 2}$ regularization term in error function is difficulties to convergence analysis, so we proposed to smoothing the usual one at the origin, resulting in the following error function with a smoothing $L_{1\backslash 2}$ regularization penalty term:

$$E(\omega) = \frac{1}{2} \sum_{j=1}^{J} \sum_{i=1}^{n} \left(o^j - g(\omega_i \cdot \xi^j) \right)^2 + \lambda \sum_{i=1}^{n} \sum_{k=1}^{p} f(\omega_{ik})^{1/2}$$

$$= \sum_{j=1}^{J} \sum_{i=1}^{n} g_{ji}(\omega_i \cdot \xi^j) + \lambda \sum_{i=1}^{n} \sum_{k=1}^{p} f(\omega_{ik})^{1/2} \tag{7}$$

In order to approximate the non-smooth function $|x|$. For definiteness and simplicity, we use the smoothing function $f(x)$ defined by:

$$f(x) = \begin{cases} -x, & x \leq -a; \\ -\dfrac{1}{8a^3} x^4 + \dfrac{3}{4a} x^2 + \dfrac{3}{8}, & -a < x < a; \\ x, & x \geq a; \end{cases} \tag{8}$$

Where a is a small positive constant. Then we have

$$f'(x) = \begin{cases} -1, & x \leq -a; \\ -\dfrac{1}{2a^3} x^3 + \dfrac{3}{2a} x, & -a < x < a; \\ 1, & x \geq a; \end{cases}$$

$$f''(x) = \begin{cases} 0, & x \leq -a; \\ -\dfrac{3}{2a^3} x^2 + \dfrac{3}{2a}, & -a < x < a; \\ 0, & x \geq a; \end{cases}$$

It is easy to get

$$f(x) \in \left[\frac{3}{8} a, +\infty\right), \qquad f'^{(x)} \in [-1, 1], \qquad f''^{(x)} \in \left[0, \frac{3}{2a}\right]$$

Where $g_j(t) = \frac{1}{2}(o^j - g(t))^2$ and $\lambda > 0$ is a penalty parameter. Then gradient function with respect to $\omega_{ik}(\, i = 1,2, \ldots, n;\ k = 1,2, \ldots, p)$.

$$E_{\omega_{ik}}(\omega) = \sum_{j=1}^{J} g'_{ji}(\omega_i \cdot \xi^j) \xi_k^j + \lambda \frac{f'(\omega_{ik})}{2 f(\omega_{ik})^{1/2}} \tag{9}$$

Let ω^0 be arbitrary initial weight. The offline gradient method with smoothing $L_{1/2}$ regularization term updates the weights $\{\omega^m\}$ iteratively by

$$\omega_{ik}^{mJ+j} = \omega_{ik}^{mJ+j-1} - \Delta_{\mathrm{j}}^{\mathrm{m}} \omega_{ik}^{mJ+j-1}, \qquad m = 0,1,2, \ldots \tag{10}$$

Where

$$\Delta_j^m \omega_{ik}^{mJ+j-1} = -\eta_m \nabla_{\omega_{ik}} E\left(\omega_{ik}^{mJ+j-1}\right) = -\eta_m E_{\omega_{ik}}\left(\omega_{ik}^{mJ+j-1}\right)$$

$$= -\eta_m \left(g'_{ji}\left(\omega_i^{mJ+j-1} \cdot \xi^j\right)\xi_k^j + \lambda \, \frac{f'\left(\omega_{ik}^{mJ+j-1}\right)}{2Jf\left(\omega_{ik}^{mJ+j-1}\right)^{1/2}} \right) \tag{11}$$

Here, $\nabla_{\omega_{ik}} E(\omega^{mJ+j-1})$ is the gradient of $E(\omega^{mJ+j-1})$ with respect to ω_{ik} $(i = 1,2,\ldots,n;\ k = 1,2,\ldots,p)$ and $\eta_m > 0$ is the learning rate.

3 MAIN RESULTS

Our assumptions in this paper are described below:

Assumption (A1): $\left|g^{(k)}(t)\right| < C$, $\left|g_{ji}^{(k)}(t)\right| < C$ $(k = 0,1,2)$ are uniformly bounded for $t \in \mathbb{R}$.

Assumption (A2): λ and η are chosen to satisfy $0 < \eta < \frac{1}{\lambda + C_1}$.

For simplicity, we denote

$$C_1 = \frac{1}{2} JCC_2^2, \qquad \mathrm{C}_2 = \max_{1\le j\le J}\left\|\xi^j\right\|,$$

Assumption (A3): the set $\Omega_0 \in \{\omega \in \Omega: E_\omega(\omega) = 0\}$ Contains finite points, where Ω is closed bounded region such that $\{\omega^m\} \subset \Omega$.

The following Lemma is a crucial tool for our analysis.

Lemma 1. Let $F\colon \Phi \subset \mathbb{R}^p \to \mathbb{R}$ $(\mathrm{p} \ge 1)$ be continuous for a bounded closed region Φ. if the set $\Phi_0 = \{x \in \Phi\colon \mathrm{F_x(x)} = 0\}$ has finite points and the sequence $\{x_n\} \in \Phi$ satisfy: $\lim_{n\to\infty}\|F_x(x_n)\| = 0$ and $\lim_{n\to\infty}\|x_{n-1} - x_n\| = 0$. Then, there exists $x^* \in \Phi_0$ such that $\lim_{n\to\infty} x_n = x^*$. Thus proof is omitted see [20].

The next theorem confirms the boundedness of the weights in the training procedure, which is a desired rewarding of adding a smoothing $L_{1\backslash 2}$ regularization penalty term.

Theorem 2. Suppose that Assumptions (A1) and (A2) are valid. That the weight sequence $\{\omega_{ik}^{mJ+j}\}$ is the generated form equ. (10). For arbitrary initial valueω^0, Then $\{\omega_{ik}^{mJ+j}\}$ $(i = 1,2,\ldots,n\,; k = 1,2,\ldots,p; m = 0,1,2,\ldots)$ are uniformly bounded, i.e., there exist positive constants $M > 0$ such that

$$\left\|\omega_{ik}^{mJ+j}\right\| \le M, \quad i = 1,2,\ldots,n\,; k = 1,2,\ldots,p; m = 0,1,2,\ldots \tag{12}$$

Proof. By applying the assumption (A1), there is a constant $A_2 > 0$ such that for all $m = 0,1,2,\ldots$,

$$\sum_{j=1}^{J}\left|g'_{ji}\left(\omega_i^{mJ+j-1} \cdot \xi^j\right)\right|\left\|\xi_k^j\right\| \le A_2 \tag{13}$$

In addition to that, for $x \in \mathbb{R}, f(x) \in \left[\frac{3}{8}a, +\infty\right), f'(x) \in [-1,1]$ holds. By the updating equs. (10) and (11), we have

$$\begin{aligned}
\left|\omega_{ik}^{mJ+j} - \omega_{ik}^{mJ+j-1}\right| &= \eta_m\left|\Delta\,\omega_{ik}^{mJ+j-1}\right| \\
&\le \eta_m\left(\left| g'_{ji}\left(\omega_i^{mJ+j-1} \cdot \xi^j\right)\xi_k^j + \lambda\, \frac{f'\left(\omega_{ik}^{mJ+j-1}\right)}{2Jf\left(\omega_{ik}^{mJ+j-1}\right)^{1/2}}\right|\right) \\
&\le \eta_m\left(A_2 + \frac{\lambda}{3a}\sqrt{6a}\right) \le \eta_m A_1
\end{aligned}$$

Where $A_1 = A_2 + (\lambda/3a)\sqrt{6a}$.

Then

$$\begin{aligned}
\left|\omega_{ik}^{(m+1)J+j} - \omega_{ik}^{mJ+j}\right| &\le \left|\omega_{ik}^{(m+1)J+j} - \omega_{ik}^{(m+1)J+j-1}\right| + \left|\omega_{ik}^{(m+1)J+j-1} - \omega_{ik}^{(m+1)J+j-2}\right| \\
&\quad + \cdots + \left|\omega_{ik}^{(m+1)J+1} - \omega_{ik}^{(m+1)J}\right| + \left|\omega_{ik}^{mJ+J} - \omega_{ik}^{mJ+j-1}\right| + \left|\omega_{ik}^{mJ+j-1} - \omega_{ik}^{mJ+j-2}\right| \\
&\quad + \cdots + \left|\omega_{ik}^{mJ+j+1} - \omega_{ik}^{mJ+j}\right| \\
&\le (j\eta_{m+1} + (J-j)\eta_m)A_1
\end{aligned} \tag{14}$$

Since

$$\left|\omega_{ik}^{(m+p)J+j}-\omega_{ik}^{mJ+j}\right| \le \left|\omega_{ik}^{(m+p)J+j}-\omega_{ik}^{(m+p-1)J+j}\right| + \left|\omega_{ik}^{(m+p-1)J+j}-\omega_{ik}^{(m+p-2)J+j}\right|$$
$$+\cdots+\left|\omega_{ik}^{(m+1)J+j}-\omega_{ik}^{mJ+j}\right|$$
$$\le A_1 j(\eta_{m+p}+\eta_{m+p-1}+\cdots.+\eta_{m+1})+C_2(J-j)(\eta_{m+p-1}+\eta_{m+p-2}+\cdots.+\eta_m)$$
$$\le JA_1\varepsilon \tag{15}$$

Therefore, the weight sequence $\{\omega_{ik}^{mJ+j}\}$ is a convergence. By the properties of convergence sequence, $\{\omega_{ik}^{mJ+j}\}$ must be a bounded sequence, so is $\|\omega_{ik}^{mJ+j}\|$. namely, there exists a suitable constant $M>0$ such that

$$\|\omega_{ik}^{mJ+j}\| \le M \tag{16}$$

Where $i=1,2,\dots,n\,;k=1,2,\dots,p;\,;m=0,1,2,\dots;j=1,2,\dots,J$.

Naturally, there also exists a constant $\bar{M}\ge 0$ such that

$$\|\Delta_{k}^{m}\omega_{i}^{mJ+j-1}\| \le \bar{M} \quad ,k=1,2,\dots,J \tag{17}$$

This proof is completed.

Theorem 3. Suppose that the error function is given by equ. (7), that the weight sequence $\{\omega^m\}$ is generated by the algorithm of equ. (11) for any initial value ω^0, and Assumption (A1) and (A2) are valid. Then we have

(i) $E(\omega^{(m+1)J}) \le E(\omega^{mJ})$,

(ii) There is $E^*\ge 0$ such that $\lim_{m\to\infty}E(\omega^{mJ})=E^*$;

(iii) $\lim\limits_{m\to\infty}\|\Delta\omega_i^{mJ}\|=0,\quad \lim\limits_{m\to\infty}\|E_\omega(\omega^{mJ})\|=0.$

Moreover, if Assumption (A3) is valid, then we have the strong convergence:

(iv) There *exists* $\omega^*\in\Omega_0$ such *that* $lim_{m\to\infty}\,\omega^m=\omega^*$.

The Proof of Theorem 3. Is divided into four parts dealing with statements $(\mathrm{i}),(\mathrm{ii}),(\mathrm{iii})$ and (iv), respectively.

Proof of (i) of the theorem 3. For convenience, we use the following notations:

$$\sigma^m=\sum_{i=1}^{n}\sum_{k=1}^{p}\left(\Delta_j^m\omega_{ik}^{mJ}\right)^2 \tag{18}$$

Using the Taylor's formula we expand $g_{ji}\left(\omega_i^{(m+1)J}\cdot\xi^j\right)$ at $\omega_i^{mJ}\cdot\xi^j$ we get

$$g_{ji}\left(\omega_i^{(m+1)J}\cdot\xi^j\right)$$
$$=g_{ji}(\omega_i^{mJ}\cdot\xi^j)+g'_{ji}(\omega_i^{mJ}\cdot\xi^j)(\Delta\omega_i^{mJ})\xi^j$$
$$+\frac{1}{2}g''_{ji}(t_{m,j})\left((\Delta\omega_i^{mJ})\xi^j\right)^2 \tag{19}$$

Where $t_{m,j}$ lies in between $\omega_i^{mJ}\cdot\xi^j$ and $\omega_i^{(m+1)J}\cdot\xi^j$.

From equ. (7) and equ. (19) we get

$$E(\omega^{(m+1)J})-E(\omega^{mJ})=\sum_{j=1}^{J}\sum_{i=1}^{n}\sum_{k=1}^{p}g'_{ji}(\omega_i^{mJ}\cdot\xi^j)\,(\Delta_j^m\omega_{ik}^{mJ})\xi_k^j+\frac{1}{2}JCC_2^2\sum_{i=1}^{n}\sum_{k=1}^{p}\left(\Delta_j^m\omega_{ik}^{mJ}\right)^2$$
$$+\lambda\sum_{i=1}^{n}\sum_{k=1}^{p}\left(f\left(\omega_{ik}^{(m+1)J}\right)^{1/2}-f\left(\omega_{ik}^{mJ}\right)^{1/2}\right)$$
$$=-\left(\frac{1}{\eta_m}-C_1\right)\sum_{i=1}^{n}\sum_{k=1}^{p}\left(\Delta_j^m\omega_{ik}^{mJ}\right)^2-\lambda\sum_{i=1}^{n}\sum_{k=1}^{p}\frac{f'^{\left(\omega_{ik}^{mJ}\right)}}{2f\left(\omega_{ik}^{mJ}\right)^{1/2}}\cdot\left(\Delta_j^m\omega_{ik}^{mJ}\right)$$
$$+\lambda\sum_{i=1}^{n}\sum_{k=1}^{p}(f\left(\omega_{ik}^{(m+1)J}\right)^{1/2}-f\left(\omega_{ik}^{mJ}\right)^{1/2} \tag{20}$$

Where $C_1 = \frac{1}{2} JCC_2^2$.

By using the Lagrange mean value theorem for $f(x)$, we have

$$E\left(\omega^{(m+1)J}\right) - E\left(\omega^{mJ}\right) = -\frac{1}{\eta_m}\sum_{i=1}^{n}\sum_{k=1}^{p}\left(\Delta_j^m \omega_{ik}^{mJ}\right)^2 + \frac{\lambda}{2}\sum_{i=1}^{n}\sum_{k=1}^{p} F''\left(t_{i,k,m}\right)\left(\Delta_j^m \omega_{ik}^{mJ}\right)^2$$

$$= -\left(\frac{1}{\eta_m} - M\lambda - C_1\right)\sum_{i=1}^{n}\sum_{k=1}^{p}\left(\Delta_j^m \omega_{ik}^{mJ}\right)^2 \qquad (21)$$

Where $t_{i,k,m} \in \mathbb{R}$ is between ω_{ik}^{mJ} and $\omega_{ik}^{(m+1)J}$, $M = \frac{\sqrt{6}}{\sqrt{a^3}}$,and $F(x) \equiv \left(f(x)\right)^{\frac{1}{2}}$. Note that

$$F'(x) = \frac{f'(x)}{2\sqrt{f(x)}}$$

$$F''(x) = \frac{2f''(x).f(x) - [f'(x)]^2}{4[f(x)]^{\frac{3}{2}}} \leq \frac{f''(x)}{2\sqrt{f(x)}} \leq \frac{\sqrt{6}}{2\sqrt{a^3}}$$

By assumption (A2), we have

$$E\left(\omega^{(m+1)J}\right) \leq E\left(\omega^{mJ}\right) \quad , m = 0,1,2,\ldots, j = 1,2,\ldots, J \qquad (22)$$

This completes the proof to statement (i) of Theorem 3.

Proof of* (ii) *of Theorem 3. Since the nonnegative sequence $\{E(\omega^{mJ})\}$ is monotone and bounded below. There must be a limit value $E^* \geq 0$ such that $\lim_{m\to\infty} E(\omega^{mJ}) = E^*$. The Proof of (ii) is thus completed.

Proof of* (iii) *of Theorem 3. It follows from assumption (A2) that $\beta > 0$. Taking $\beta = \frac{1}{\eta_m} - \lambda M - C_1$ and using equ. (21), we have

$$E\left(\omega^{(m+1)J}\right) \leq E\left(\omega^{mJ}\right) - \beta\sigma^m \leq \cdots \leq E\left(\omega^0\right) - \beta\sum_{k=0}^{m}\sigma^k$$

Since $\left(\omega^{(m+1)J}\right) \geq 0$, we have

$$\beta\sum_{k=0}^{m}\sigma^k \leq E\left(\omega^0\right).$$

Let $m \to \infty$, then

$$\beta\sum_{k=0}^{\infty}\sigma^k \leq \frac{1}{\beta}E\left(\omega^0\right) < \infty.$$

these results to

$$\lim_{m\to\infty}\sigma^m = \lim_{m\to\infty}\sum_{i=1}^{n}\sum_{k=1}^{p}\left(\Delta_j^m \omega_{ik}^{mJ}\right)^2 = 0.$$

It follows from equs. (9) – (11) and equ. (19) that

$$\lim_{m\to\infty}\left\|\Delta_j^m \omega_{ik}^{mJ}\right\| = 0, \quad \lim_{m\to\infty}\left\|E_\omega\left(\omega^{mJ}\right)\right\| = 0. \qquad (23)$$

The proof to (iii) is thus completed.

Proof of* (iv) *of Theorem 3. From equ. (23) leads to: existing $E^* \in \Phi_0$ such that $\lim_{m\to\infty}(\omega^{mJ}) = \omega^*$. This completes the proof of *(iv)*.

4 CONCLUSION

Convergence results are established for the offline gradient method with smoothing $L_{1\backslash 2}$ regularization term for training multi-output feedforward neural networks. The monotonicity of the error function and weight boundedness for the offline gradient with smoothing $L_{1\backslash 2}$ regularization are presented, both weak and strong convergence results are proved, which will provides a strong theoretical support for many applications on multi-output neural networks.

ACKNOWLEDGMENT

We gratefully acknowledge Dalanj University and Dalian University of Technology for supporting this research. Special thanks to Prof. Wei Wu and Dr. Yan Liu for their kind helps during the period of the research.

REFERENCES

[1] Liang Y. C. et al, Successive Approximation training algorithm for feed forward neural networks. Neurocomputing, vol. 42: 311-322, 2002.

[2] Zhang X.S, Neural Networks in Optimization. Boston: Kluwer Academic Publishers, 2000.

[3] Jie Yang, Wenyu Yang, Wei Wu , A remark on the error back propagation learning algorithm for spiking neural networks. Applied Mathematics Letters, vol. 25: 1118–1120, 2012.

[4] Kong J. Wu W., Online gradient methods with a punishing term for neural networks. Northeast Math. J., vol. 173: 371-378, 2001.

[5] Wu W., Computation of Neural Networks. Beijing : Higher Education Press, 2003.

[6] Fine T. L. and Mukherjee S., Parameter convergence and learning curves for neural networks, Neural Computation, vol. 11: 747–769, 1999.

[7] Li Z. X., Wu. W. and Tian Y. L., Convergence of an online gradient method for feedforward neural networks with stochastic inputs, Journal of Computer Applied Mathematics, vol. 163: 165–176, 2004.

[8] Finnoff W., Diffusion approximations for the constant learning rate back propagation algorithm and rcsistance to local minima. Neural comput, vol. 6: 285-295, 1994.

[9] Hrnik K. and Kuan C. M., Convergence of learning algorithms with constant learning rates, IEEE Trans. Neural Networks, vol. 3b: 484 – 489, 1991.

[10] Hinton G., Connectionist learning procedures, Artificial intelligence, vol. 40: 185- 234, 1989.

[11] Lonne S. and Irwin G., Improving neural network training solutions using regularization, Neurocomputing, vol. 37: 71-90, 2001.

[12] Setiono R, A penalty- function approach for pruning feedforward neural networks. Neural Computation, vol. 9: 251 – 254, 1997.

[13] H. Shao, W. Wu and L. Liu, convergence and monotonicity of an online gradient method with penalty for neural networks. WSEAS Trans Math vol. 6, no. 3: 469 – 476, 2007.

[14] Wei Wu, Qinwei Fan, Jacek M. Zurada , Jin Wang, Dakun Yang and Yan Liu, Batch gradient method with smoothing $L_{1/2}$Regularization for training of feedforward neural networks, Neural Networks, vol. 50: 72-78, 2014.

[15] Qin wei Fan, Jacek M. Zurada and Wei Wu, Convergence of online gradient method for feedforward neural networks with smoothing $L_{1/2}$ regularization penalty. Neurocomputing, vol. 131: 208-216, 2014.

[16] Fine T. L. and Mukherjee S., Parameter convergence and learning curves for neural networks, Neural Computation, vol. 11: 747- 769, 1999.

[17] Hongmei Shao, Wei Wu, and Feng Li, Convergence of online gradient method with a penalty term for feedforward neural networks with stochastic inputs. Numerical Mathematics, vol. 1, no. 14: 87-96, 2005.

[18] Wei Wu, Hongmei Shao and Zhengxue Li , Convergence of batch BP algorithm with penalty for FNN training. Lecture Notes in Computer Science, vol. 4232: 562-569, 2006.

[19] Huisheng Zhang and Wei Wu, Boundedness and convergence of batch back-propagation algorithm with penalty for neural networks. Neurocomputing, vol. 89: 141-146, 2012.

[20] Yuan Y.X., and Sun W.Y., Optimization Theory and Methods. Beijing . Science Press, 2001.

Khidir Shaib Mohamed (PhD student in Computational Mathematics) received the B.Sc in Mathematics from Dalanj University – Dalanj – Sudan (2006) and M.Sc in Applied Mathematics from Jilin University – Changchun – China (2011). He work as a lecturer of mathematics at College of Science – Dalanj University, Dalanj – Sudan since (2011). Now he is a PhD student in Applied Mathematics at School of Mathematical Sciences, Dalian University of Technology, Dalian – China, since 2012. He is very interest for future theoretical researches for the analysis and improvement of learning algorithms in neural networks (E-mail: khshm7@yahoo.com).

Yousif Shoaib Mohammed (Assistant Professor of Computational Physics) received the B.Sc in Physics from Khartoum University – Oudurman – Sudan (1994) and High Diploma in Solar Physics from Sudan University of Science and Technology – Khartoum – Sudan (1997) and M.Sc in Computational Physics (Solid State – Magnetism) from Jordan University – Amman – Jordan and PhD in Computational Physics (Solid State – Magnetism – Semi Conductors) from Jilin University – Changchun – China (2010). He worked at Dalanj University since 1994 up to 2013 and worked as Researcher at Africa City of Technology – Khartoum – Sudan since 2012. Then from 2013 up to now at Qassim University – Kingdom of Saudi Arabia (E-mail: yshm@yahoo.com).