Boundedness and Convergence of Batch Gradient Method for Training Pi-Sigma Neural Network with Inner-Penalty and Momentum

Kh. Sh. Mohamed, Xiong Yan, Zhengxue Li, Z. A. Habtamu, Abdrhaman. M. Adam

***Abstract*— In the process industries convergence of a batch gradient method with inner-penalty and adaptive momentum is inspection for training pi-sigma neural networks. The role of the usual penalty is considered, which is a term proportional to the norm of the weights to control the magnitude of the weights and improve the generalization performance of the network. The monotonicity theorem and two convergence theorems of our gradient algorithm with inner-penalty term is guaranteed during the training iteration.**

***Index Terms*— Convergence, pi-sigma neural network, batch gradient method, inner-penalty, momentum, boundedness**

I. INTRODUCTION

The traditional conventionally artificial neural networks (ANN) compared together with higher order neural networks (HONN), the two models have been used with different architecture and learning rules have become popular tool to solve wide range of problems like classification, association, recognition and control. Thus, in [2,4,7,9] HONN models have shown superior performance than traditional ANN in [3,8,10,12,18] on forecasting, classification and regression problems because the HONN have several unique characteristics, including such that (greater storage capacity, stronger approximation property, higher fault tolerance capability, faster convergence, ...). The pi-sigma network (PSN), which is a class of HONN shown by Shin and Ghosh [1]. This network combines the fast training algorithm abilities of single layered networks with the non-linear mapping of HONN and utilizes product cells as the output units to indirectly the capabilities of higher-order networks while using a fewer numbers diminution of weights and procession units and have been used effectively in pattern classification and approximation. The regularization (penalty) method is often append into the learning process and have prove to be efficient to improve the generalization capability and to the magnitude of the network weights [6,11,15,20]. Especially in [16] the penalty term celled inner-penalty and it is useful to prove capability and magnitude network training.

Revised Version Manuscript Received on January 04, 2016.

Kh. Sh. Mohamed, School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, PR China.

Xiong Yan, School of Science, Liaoning University of Science & Technology, Anshan 114051, China.

Zhengxue Li, School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, PR China.

Z. A. Habtamu, School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, PR China.

Abdrhaman M. Adam, School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, PR China.

To speed up and constancy the training iteration procedure, a momentum term is often insert to the increment formula for the weights so that the new weight updating rule becomes a combination of the present gradient of the error function and the previous weight updating increment [13,14,17,19,21]. In recent years the online and batch gradient method with momentum has been widely used under the assumption that the error function is quadratic [23-25]. The new error function with penalty and is decreasing monotonically and the batch gradient method with both penalty and momentum is deterministically convergent under the momentum coefficient and penalty parameter are both is positive constant. For related work we mention [5,22] where a feddforward network is considered for two-three layers cases.

The rest of this paper is organized as follows. In section II, the gradient algorithm with inner-penalty and momentum is presented for training pi-sigma neural network. In section III, the main convergence result are presented. The rigorous proofs of the main results are provided in Section IV. Finally, some conclusions are drawn in Section V.

In this paper, the notation $\|\cdot\|$ denotes the Euclidean vector norm.

II. BATCH GRADIENT ALGORITHM WITH INNER-PENALTY AND MOMENTUM TERM

For a given set of training examples $\{\xi^j, y^j\}_{j=1}^J \subset \mathbb{R}^p \times \mathbb{R}^J$, J is the numbers of training examples. Let us describe the structure of neural network, which suppose that of an input layer, summation layer and product layer are $P, N,$ and 1 respectively. Let $g: \mathbb{R} \to \mathbb{R}$ be a given activation function for the output layer, which is often, but not necessarily, selected as the logistic function $g(x) = 1/(1+e^{-x})$. We denote by $w_i = (w_{i1}, \dots, w_{ip})^T \in \mathbb{R}^p (1 \le i \le N)$ the weight vectors connecting the input and summing units, and write $w = (w_1^T, \dots, w_N^T)^T \in \mathbb{R}^{Np}$. Here we have added a special input unit $\xi_p = -1$ corresponding to the biases $w_{jp} (1 \le j \le N)$. Note that the weights from summing units to product unit are fixed to 1. For any given input $\xi \in \mathbb{R}^p$ the output of the neural network by $y^j = g(\prod_{k=1}^N (w_k \cdot \xi^j))$. Our error function with inner-penalty term take the form:

$$E(W) = \sum_{j=1}^{J} g_j \left(\prod_{k=1}^{N} (w_k \cdot \xi^j) \right) + \frac{\lambda}{2} \sum_{j=1}^{J} \sum_{i=1}^{N} (w_i \cdot \xi^j)^2 \quad (1)$$

The structure of pi-sigma neural network is shown in fig. 1. Below

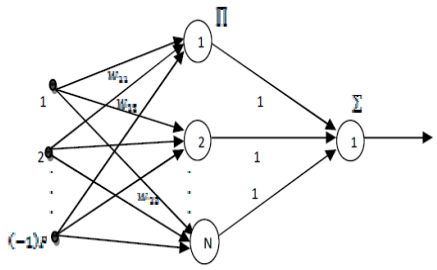


Fig. 1. A Pi-sigma network

where $\lambda > 0$ is penalty coefficient and $g_j(t) = \frac{1}{2}\left(o^j - g(t)\right)^2$ $(j = 1,..,J, t \in \mathbb{R})$. Then gradient of the error function with respect to Ws given by:

$$E_{w_i}(W) = \sum_{j=1}^{J} g_j^{'}\left(\prod_{k=1}^{N}(w_k \cdot \xi^j)\right)\prod_{\substack{k=1\\k\neq i}}^{N}(w_k \cdot \xi^j)\xi^j + \lambda(w_i \cdot \xi^j)\,\xi^j \quad (2)$$

Now we introduce the batch gradient algorithm with inner-penalty term and momentum (BGPM). Let $\{\xi^{n1}, \dots, \xi^{nJ}\}$ be a stochastic permutation of $\{\xi^1, \dots, \xi^J\}$ in the *n-th* cycle of the training iteration. Given initial weights w^0, we proceed to refine them iteratively by

$$W^{nJ+j} = W^{nJ+j-1} + \Delta_j^n W^{nJ+j} \quad (3)$$

with

$$\Delta_j^n w_i^{nJ+j} = -\eta_n E_{w_i^{nJ+j-1}}(W) + \alpha_i^{n,j}\,\Delta_j^n w_i^{nJ+j-1} \quad (4)$$

where $n = 0,1,\dots, j = 1,2,\dots,J$ and $\eta_n > 0$ is the learning rate in the *n-th* training cycle, $\alpha_i^{n,j}\Delta_j^n \omega_i^{nJ+j-1}$ is the so-called momentum term and $\alpha_i^{n,j}$ is the momentum coefficient. For the sake of description, we denote

$$p_i^{n,t,j} = E_{w_i^{n+t-1}}(W) \quad (5)$$

Particularly when $t = 1$ denote

$$p_i^{n,1,j} = p_i^{n,j} = E_{w_i^{nJ}}(W) \quad (6)$$

Then there holds

$$E_{w_i^{nJ}}(W) = \sum_{j=1}^{J} p_i^{n,j} \quad (7)$$

and the learning rule (4) becomes

$$\Delta_j^n w_i^{nJ+j} = -\eta_n p_i^{n,j,j} + \alpha_i^{n,j}\,\Delta_j^n w_i^{nJ+j-1} \quad (8)$$

In this work, by choosing an initial $\eta_0 \in (0,1]$ and positive constant β, we inductively determine η_n in (8) by (cf.[25])

$$\frac{1}{\eta_{n+1}} = \frac{1}{\eta_n} + \beta, \quad n = 0,1,2\dots \quad (9)$$

It is easy to get from (9) that $\eta_n = -\eta_0/(1 + n\beta\eta_0)$ for $n = 0,1,\dots$ hence there hold $\eta_n = o(1\backslash n)$ and for $\eta_n \to 0\ as\ n \to \infty$ and for the momentum coefficients $\alpha_i^{n,j}$ in (8), then we choose them by the rule

$$\alpha_i^{n,j} = \begin{cases} \dfrac{\eta_n^2 \|p_i^{n,j}\|}{\|\Delta_j^n w_i^{nJ+j-1}\|} & if \|\Delta_j^n w_i^{nJ+j-1}\| \neq 0 \\ 0 & else \end{cases} \quad (10)$$

III. MAIN RESULTS

The following conditions will be used in this paper

(A1) $|g_j(t)|, |g_j^{'}(t)| \& |g_j^{''}(t)| \le C, \forall t \in \mathbb{R}, \forall 1 \le j \le J.$

(A2) $max_{1\le j\le J}\left\{\left\|\xi^j\right\|, \left\|\xi^j\right\|^2\right\} \& \left|w_i^k \cdot \xi^j\right| \le C, \forall 1 \le j \le J, 1 \le i \le N, k = 0,1,\dots$

(A3) inequality (50) is valid, and β and η_0 in (9) satisfy: $\beta > max\{1, \tilde{\beta}\}$ and $\eta_0 \le min\{1, 1/\tilde{\beta} - 1/\beta\}$.

Set

$$\tilde{\beta} = C_{12} + C_{13} \quad (11)$$

(A4) The set $\Omega_0 \in \{w \in \Omega : E_w(W) = 0\}$ contains finite points, where Ω is closed bounded region such that $\{w^m\} \subset \Omega$.

Theorem 1 Suppose that the error function $E(W)$ be given by (1), let Assumptions (A1) and (A2) be satisfied, and let the weight $\{W^k\}$ be generated by the algorithm (4). Then there hold

$$E\left(W^{(n+1)J}\right) \le E(W^{nJ}), \quad n = 0,1,\cdots$$

Theorem 2 Under the same Assumption of Theorem 1, the weight sequence $\{W^k\}$ generated by (4) is uniformly bounded.

Theorem 3 Suppose that the error function $E(W)$ be defined in (1) and the learning rate $\{\eta_n\}$ be determined by (9). Given any initial values w^0, the weights $\{W^k\}$ are generated by the algorithm (4). If Assumptions (A1) - (A3) are valid, there holds the following weak convergence result:

$$\lim_{k\to\infty} \|E_w(W^k)\| = 0.$$

Furthermore, if Assumption (A4) is also valid, there holds the following strong convergence result: There exists $w^* \in \Omega$ such that

$$\lim_{k\to\infty} W^k = W^*.$$

IV. PROOFS

For convergence notation, we denote

$$\psi_\iota^{nJ+j} = \prod_{k=1}^{N}(w_k^{nJ+j} \cdot \xi^j), 1 \le \iota \le J, 1 \le j \le J \quad (12)$$

$$\varphi_{\iota,i}^{nJ+j} = \prod_{\substack{k=1\\k\neq i}}^{N}(w_k^{nJ+j} \cdot \xi^j), 1 \le \iota \le J, 1 \le j \le J \quad (13)$$

$$r_i^{n,j} = p_i^{n,j,j} - p_i^{n,j}, \quad n = 0,1,2,\dots, j = 1,2,\dots,J. \quad (14)$$

$$d_i^{n,j} = w_i^{(n+1)J} - w_i^{nJ}, \quad n = 0,1,2,\dots \quad (15)$$

The following Lemmas are useful in proof of our convergence results.

Lemma 2 Suppose that the $\{\eta_n\}$ be given by (9). There hold

$$(i)\ \ 0 < \eta_n < \eta_{n+1} \le 1, \quad n = 1,2,\dots \quad (16)$$

$$(ii)\ \frac{\tau}{n} < \eta_n < \frac{\rho}{n}, \ \tau = \frac{\eta_0}{1+\eta_0\beta}, \ \rho = \frac{1}{\beta}, \ n = 1,2,\dots \quad (17)$$

Proof This Lemma is easy to validate by virtue of (10) and $\eta_0 \in (0,1]$.

The next Lemmas estimate $r_i^{n,j}$ and the

change of error function.

Lemma 3 Suppose that Assumption (A1) is satisfied, there exist a constants $C_1, C_2, C_3 > 0$, such that for any $n = 1,2, \ldots$

$$(i)\ \sum_{j=1}^{J}\|r_i^{n,j}\| \le C_1\eta_n\sum_{k=1}^{N}\sum_{t=1}^{J}\|p_k^{n,t}\| \tag{18}$$

$$(ii)\ \|d_i^{n,j}\| \le \sum_{k=1}^{N}\left\|\sum_{t=1}^{J}p_k^{n,t}\right\| + C_2\eta_n^2\sum_{k=1}^{N}\sum_{t=1}^{J}\|p_k^{n,t}\| \tag{19}$$

$$(iii)\ \|d_i^{n,j}\|^2 \le C_3\eta_n^2\sum_{k=1}^{N}\sum_{t=1}^{J}\|p_k^{n,t}\|^2 \tag{20}$$

Proof By (8), we have

$$w_i^{nJ+j-1} - w_i^{nJ} = \sum_{j=1}^{J}\left(\alpha_i^{n,j}\Delta_j^n w_i^{nJ+j-1} - \eta_n p_i^{n,j}\right) \tag{21}$$

This together with (9), (14) , $0 < \eta_n \le 1$ gives

$$\begin{aligned}
&\| w_i^{nJ+j-1} - w_i^{nJ} \|\\
&\le \sum_{j=1}^{J}\left(\alpha_i^{n,j}\|\Delta_j^n w_i^{nJ+j-1}\| + \eta_n\|\eta_n p_i^{n,j}\| + \eta_n\|r_i^{n,j}\|\right)\\
&\le \sum_{j=1}^{j}\left(\eta_n^2\|p_i^{n,j}\| + \eta_n\|p_i^{n,j}\| + \eta_n\|r_i^{n,j}\|\right)\\
&\le \eta_n\sum_{i=1}^{N}\left(\sum_{t=1}^{J}2\|p_k^{n,t}\| + \sum_{t=1}^{J}\|r_k^{n.t}\|\right)
\end{aligned} \tag{22}$$

Using Assumption (A2), (21) and Cauchy-Schwartz, we have

$$\begin{aligned}
&|\psi_\iota^{nJ+j-1} - \psi_\iota^{nJ}|\\
&\le \left|\prod_{k=1}^{N-1}\left(w_k^{nJ+j}.\xi^j\right)\right|\left|\left(w_N^{nJ+j-1} - w_N^{nJ}\right)\xi^j\right|\\
&+ \left|\prod_{k=1}^{N-2}\left(w_k^{nJ+j-1}.\xi^j\right)\left(w_N^{nJ-1}.\xi^j\right)\right|\left|\left(w_{N-1}^{nJ+j-1} - w_{N-1}^{nJ}\right)\xi^j\right|\\
&+ \cdots + \left|\prod_{k=1}^{N}\left(w_k^{nJ}.\xi^j\right)\right|\left|\left(w_1^{nJ+j-1} - w_1^{nJ}\right)\xi^j\right|\\
&\le C^{N-1}\|\xi^j\|\sum_{k=1}^{N}\left\|\sum_{t=1}^{J}\left(2p_k^{n,t} + r_k^{n,t}\right)\right\|\\
&\le C_4\sum_{k=1}^{N}\left(\sum_{t=1}^{J}2\|p_k^{n,t}\| + \sum_{t=1}^{J}\|r_k^{n,t}\|\right)
\end{aligned} \tag{23}$$

where $C_4 = C^N (1 \le t \le J, 1 \le i \le N, n = 0,1,2, \cdots)$.

Similarly, easy to get

$$\begin{aligned}
&|\varphi_\iota^{nJ+j-1} - \varphi_\iota^{nJ}|\\
&\le \acute{C}_4\sum_{k=1}^{N}\left(\sum_{t=1}^{J}2\|p_k^{n,t}\| + \sum_{t=1}^{J}\|r_k^{n,t}\|\right)
\end{aligned} \tag{24}$$

where $\acute{C}_4 = C^{N-1} (1 \le t \le J, 1 \le i \le N, n = 0,1,2, \cdots)$.

By Assumptions (A1), (A2), (21)- (24) and Mean Value Theorem, we have

$$\begin{aligned}
\|r_i^{n,j}\| &= \left|g_j'\left(\psi_\iota^{nJ+j-1}\right)\varphi_{\iota,i}^{nJ+j-1}\xi^{nj} - g_j'\left(\psi_\iota^{nJ}\right)\varphi_{\iota,i}^{nJ}\xi^{nj}\right|\\
&\quad + \lambda\left\|w_i^{nJ+j-1} - w_i^{nJ}\right\|\|\xi^{nj}\|^2\\
&= \left|g_j''(t_1)\left(\varphi_{\iota,i}^{nJ+j-1}\right)\left(\psi_\iota^{nJ+j-1} - \psi_\iota^{mJ}\right)\right|\|\xi^{nj}\|\\
&\quad + \left\|g_j'\left(\psi_\iota^{nJ}\right)\left(\varphi_{\iota,i}^{nJ+j-1} - \varphi_{\iota,i}^{nJ}\right)\right\|\|\xi^{nj}\|\\
&\quad + \lambda\left\|w_i^{nJ+j-1} - w_i^{nJ}\right\|\|\xi^{nj}\|^2\\
&\le C_5\eta_n\sum_{k=1}^{N}\left(\sum_{t=1}^{J}2\|p_k^{n,t}\| + \sum_{t=1}^{J}\|r_k^{n,t}\|\right)
\end{aligned} \tag{25}$$

where $t_1 \in \mathbb{R}$ line segment between $w_k^{(n+1)J}\cdot\xi^{nj}$ and $w_k^{nJ}\cdot\xi^{nj}$ and $C_5 = \left(C_4C^{N+1} + \acute{C}_4C^2 + \lambda C\right)$. Not that for denotation functions (6) and (14) imply

$$r_i^{n,1} = 0 \tag{26}$$

This together with (25) hold

$$\begin{aligned}
\|r_i^{n,2}\| &\le C_5\eta_n\sum_{i=1}^{N}\left(2\|p_i^{n,1}\| + \|r_i^{n,1}\|\right)\\
&\le 2C_5\eta_n\sum_{i=1}^{N}\|p_i^{n,1}\|
\end{aligned} \tag{27}$$

and

$$\begin{aligned}
\|r_i^{n,3}\| &\le C_5\eta_n\sum_{i=1}^{N}\left(2\|p_i^{n,2}\| + \|r_i^{n,2}\|\right)\\
&\le 2C_5(1 + C_5)\eta_n\sum_{i=1}^{N}\left(\|p_i^{n,2}\| + \|p_i^{n,1}\|\right)
\end{aligned} \tag{28}$$

Applying an induction on $\|r_i^{n,j}\|$, we have for $2 \le j \le J$

$$\|r_i^{n,j}\| \le 2C_5(1 + C_5)^{j-2}\eta_n\sum_{k=1}^{N}\sum_{t=1}^{J}\|p_k^{n,t}\| \tag{29}$$

A sum of $j = 1,2, \ldots, J$ yields Lemma $3(i)$. Immediately :

$$\sum_{j=1}^{J}\|r_i^{n,j}\| = \sum_{j=2}^{J}\|r_i^{n,j}\| \le C_1\eta_n\sum_{k=1}^{N}\sum_{t=1}^{J}\|p_k^{n,t}\| \tag{30}$$

where $C_1 = 2C_5\sum_{j=1}^{J}(1 + C_5)^{j-2}$. Next, we prove Lemma $3(ii)$. In view of (15) and (21), we get

$$d_i^{n,j} = \sum_{j=1}^{J}\left(\alpha_i^{n,j}\Delta_j^n w_i^{nJ+j-1} - \eta_n p_i^{n,j} - \eta_n r_i^{n,j}\right) \tag{31}$$

Setting $C_2 = (1 + C_1)$ and using (15) and (30), there holds

$$\begin{aligned}
\|d_i^{n,j}\| &\le \eta_n\sum_{k=1}^{N}\left\|\sum_{t=1}^{J}p_k^{n,t}\right\| + \eta_n\sum_{k=1}^{N}\sum_{t=1}^{J}\|r_k^{n,t}\|\\
&\quad + \sum_{k=1}^{N}\sum_{t=1}^{J}\alpha_i^{n,t}\|\Delta_t^n w_k^{nJ+t-1}\|\\
&\le \eta_n\sum_{k=1}^{N}\left\|\sum_{t=1}^{J}p_k^{n,t}\right\| + C_2\eta_n^2\sum_{k=1}^{N}\sum_{t=1}^{J}\|p_k^{n,t}\|\\
&\quad + \eta_n^2\sum_{k=1}^{N}\sum_{t=1}^{J}\|p_k^{n,t}\|\\
&\le \eta_n\sum_{k=1}^{N}\left\|\sum_{t=1}^{J}p_k^{n,t}\right\| + C_3\eta_n^2\sum_{k=1}^{N}\sum_{t=1}^{J}\|p_k^{n,t}\|
\end{aligned} \tag{32}$$

Finally, we prove Lemma $3(iii)$ by virtue of (32). Again using $0 < \eta_n \le 1$, the estimation (32) can be rewritten as

$$\left\|d_i^{n,j}\right\| \le (1+C_2)\ \eta_n \sum_{k=1}^{N}\sum_{t=1}^{J}\left\|p_k^{n,t}\right\| \tag{33}$$

Squaring two sides of (33) and applying Cauchy-Schwartz inequality, we have

$$\left\|d_i^{n,j}\right\|^2 \le (1+C_2)^2\eta_n^2\left(\sum_{k=1}^{N}\sum_{t=1}^{J}\left\|p_k^{n,t}\right\|\right)^2 \le C_3\eta_n^2\sum_{k=1}^{N}\sum_{t=1}^{J}\left\|p_k^{n,t}\right\|^2 \tag{34}$$

where $C_3 = J(1+C_2)^2$. The proof it is completed.

Lemma 4 If Assumption (A1) is valid and η_n satisfies (9), there hold

$$\text{(i)}\left|\sum_{i=1}^{N}-\eta_n\left(\sum_{j=1}^{J}\left(p_i^{n,j}\cdot r_i^{n,j}\right)\right)\right| \le C_6\eta_n^2\sum_{k=1}^{N}\sum_{t=1}^{J}\left\|p_k^{n,t}\right\|^2 \tag{35}$$

$$(ii)\left|\sum_{i=1}^{N}\sum_{j=1}^{J}\left(p_i^{n,j}\cdot\alpha_i^{n,j}\,\Delta_j^n w_i^{nJ+j-1}\right)\right| \le J\eta_n^2\sum_{k=1}^{N}\sum_{t=1}^{J}\left\|p_k^{n,t}\right\|^2 \tag{36}$$

Proof. It is similar to the proof of Lemma 3 in [26] and thus omitted.

Lemma 5 There is a positive constant γ independent of n such that

$$E\left(W^{(n+1)J}\right) - E\left(W^{nJ}\right) \le -\eta_n\sum_{k=1}^{N}\left\|\sum_{t=1}^{J}p_k^{n,t}\right\|^2 + \gamma\eta_n^2\sum_{k=1}^{N}\sum_{t=1}^{J}\left\|p_k^{n,t}\right\|^2 \tag{37}$$

Proof Let $\{\xi^{n1},\dots,\xi^{nJ}\}$ be a permutation of $\{\xi^1,\dots,\xi^J\}$ in the *n-th* cycle of training iteration. Let $\xi^{(n+1)j} = \xi^{nh_j}$ $(j = 1,2,\dots,J)$, where $\{h_1,\dots,h_J\}$ is a stochastic permutation of the subscript set $(j = 1,2,\dots,J)$. In view of (2) and (4), there holds that for $i = 1,2,\dots,N$

$$E_w(W^{nJ}) = \sum_{j=1}^{J}p_i^{n,h_j} = \sum_{j=1}^{J}p_i^{n,j} \quad i = 0,1,\dots,N \tag{38}$$

By the Taylor Expansion, (31), (38) , we have

$$\mathrm{g}_j\left(\psi_i^{(n+1)J}\right) - \mathrm{g}_j\left(\psi_i^{nJ}\right) = \mathrm{g}_j'\left(\psi_i^{nJ}\right)\sum_{i=1}^{N}\left(\psi_i^{(n+1)J}\right)\left(d_i^{n,j}\right)\xi^{nh_j} + \frac{1}{2}\mathrm{g}''(t_2)\left(\psi_i^{(n+1)J}-\psi_i^{nJ}\right)^2 + \frac{1}{2}\sum_{\substack{i_1,i_2=1\\ i_1\neq i_2}}^{N}\left(\prod_{\substack{k=1\\ k\neq i_1,i_2}}^{N}t_3\right)\cdot\left(d_{i_1}^{n,j}\cdot d_{i_2}^{n,j}\right)\cdot\left(\xi^{nh_j}\right)^2 \tag{39}$$

where $t_2, t_3 \mathbb{R}$ is a vector between $\prod_{k=1}^{N}(w_k^{(n+1)J}\cdot\xi^j)$ and $\prod_{k=1}^{N}(w_k^{nJ}\cdot\xi^j)$. By the Taylor Expansion, (2), (4), (31), (38) and (39) , we get

$$\begin{aligned}
&E\left(W^{(n+1)J}\right)\\
&=\sum_{j=1}^{J}\mathrm{g}_{(n+1)j}\left(\psi_i^{(n+1)J}\right)+\frac{\lambda}{2}\sum_{i=1}^{N}\sum_{j=1}^{J}\left(w_i^{(n+1)J}\cdot\xi^{(n+1)j}\right)^2\\
&=\sum_{j=1}^{J}\mathrm{g}_{nh_j}\left(\psi_i^{nJ}\right)+\frac{\lambda}{2}\sum_{i=1}^{N}\sum_{j=1}^{J}\left(w_i^{(n+1)J}\cdot\xi^{nh_j}\right)^2\\
&=\sum_{j=1}^{J}\mathrm{g}_{nh_j}\left(\psi_i^{nJ}\right)+\frac{\lambda}{2}\sum_{i=1}^{N}\sum_{j=1}^{J}\left(w_i^{nJ}\cdot\xi^{nh_j}\right)^2\\
&+\sum_{j=1}^{J}\left(\mathrm{g}_{nh_j}'\left(\psi_i^{nJ}\right)\left(\varphi_{i,i}^{nJ}\right)+\lambda\left(w_i^{nJ}\cdot\xi^{nh_j}\right)\right)\cdot\left(d_i^{n,j}\cdot\xi^{nh_j}\right)\\
&+\frac{1}{2}\sum_{j=1}^{J}\mathrm{g}''(t_2)\left(\left(\psi_i^{(n+1)J}\right)-\left(\psi_i^{nJ}\right)\right)^2+\frac{\lambda}{2}\sum_{j=1}^{J}\left(d_i^{n,j}\cdot\xi^{nh_j}\right)^2\\
&+\frac{1}{2}\sum_{j=1}^{J}\mathrm{g}_{nh_j}'\left(\psi_i^{nJ}\right)\sum_{\substack{i_1,i_2=1\\ i_1\neq i_2}}^{N}\left(\prod_{\substack{k=1\\ k\neq i_1,i_2}}^{N}t_3\right)\left(d_{i_1}^{n,j}\cdot d_{i_2}^{n,j}\right)\left(\xi^{nh_j}\right)^2\\
&\le E\left(W^{nJ}\right)+\sum_{j=1}^{J}\left(p_i^{n,h_j}\cdot d_i^{n,j}\right)+\frac{1}{2}C_7\sum_{k=1}^{N}\sum_{t=1}^{J}\left\|p_k^{n,t}\right\|^2\\
&\le E\left(W^{nJ}\right)-\eta_n\sum_{k=1}^{N}\left\|\sum_{t=1}^{J}p_k^{n,t}\right\|^2\\
&+\frac{1}{2}C_7\,\eta_n\sum_{k=1}^{N}\sum_{t=1}^{J}\left\|p_k^{n,t}\right\|^2+\Delta_m
\end{aligned} \tag{40}$$

where $t_3 \in \mathbb{R}$ is a vector between $w_k^{(n+1)J}\cdot\xi^j)$ and $w_k^{nJ}\cdot\xi^j$, $C_7 = \frac{C}{2}(C_4^2 + \lambda JC + C^2C_3)$ and

$$\Delta_m = \sum_{i=1}^{N}\left(-\eta_n\sum_{j=1}^{J}\left(p_i^{n,j}\cdot r_i^{n,j}\right)\right) + \sum_{i=1}^{N}\left(\sum_{j=1}^{J}\left(p_i^{n,j}\cdot\alpha_i^{n,j}\ \Delta_j^n w_i^{nJ+j-1}\right)\right)$$

Then we have

$$E\left(W^{(n+1)J}\right) - E\left(W^{nJ}\right) \le -\eta_n\sum_{k=1}^{N}\left\|\sum_{t=1}^{J}p_k^{n,t}\right\|^2 + (J+C_7+C_8)\eta_n^2\sum_{k=1}^{N}\sum_{t=1}^{J}\left\|p_k^{n,t}\right\|^2 \tag{41}$$

Set

$$\gamma = J + C_6 + C_7 \tag{42}$$

Obviously, γ is a positive constant independent of the iteration *n*. Immediately and finish the proof of this Lemma.

The next Lemma is also a critical step to the proof a monotonicity of the error $\{E(W^{nJ})\}$

Lemma 6 Let $\{\eta_n\}$ be given by (10) and, if Assumption (A1) - (A3) are valid, there holds

$$\sum_{k=1}^{N}\left\|\sum_{t=1}^{J}p_k^{n,t}\right\|^2 \ge \gamma\eta_n\sum_{k=1}^{N}\sum_{t=1}^{J}\left\|p_k^{n,t}\right\|^2, \tag{43}$$

then

$$\sum_{k=1}^{N}\left\|\sum_{t=1}^{J}p_k^{n+1,t}\right\|^2 \geq \gamma\eta_{n+1}\sum_{k=1}^{N}\sum_{t=1}^{J}\|p_k^{n+1,t}\|^2 \quad (44)$$

Proof By (13), noting that $p_i^{n,h_j} = p_i^{n,j}$ and the mean value theorem and, we get

$$p_i^{n+1,j} = p_i^{n,j} + r_i^{n,j} \quad (45)$$

Applying the triangle inequality to (45) and using (25) and (33), we get

$$\|p_i^{n+1,j}\| \leq \|p_i^{n,j}\| + C_5\eta_n\sum_{k=1}^{N}\left(\sum_{t=1}^{J}2\|p_k^{n,t}\| + \sum_{t=1}^{J}\|r_k^{n,t,t}\|\right)$$
$$\leq \|p_i^{n,j}\| + C_8\eta_n\sum_{k=1}^{N}\sum_{t=1}^{J}\|p_k^{n,t}\| \quad (46)$$

where $C_8 = C_5(1 + C_1)$. Thus

$$\sum_{k=1}^{N}\sum_{t=1}^{J}\|p_k^{n+1,t}\|^2 \leq \sum_{k=1}^{N}\sum_{t=1}^{J}\|p_k^{n,t}\|^2$$
$$+2C_8\eta_n\sum_{k=1}^{N}\left(\sum_{t=1}^{J}\|p_k^{n,t}\|\right)^2 + C_8^2\eta_n^2\sum_{k=1}^{N}\left(\sum_{t=1}^{J}\|p_k^{n,t}\|\right)^2$$
$$\leq \left(1 + C_9\eta_n(1+\eta_n)\right)\sum_{k=1}^{N}\sum_{t=1}^{J}\|p_i^{n,t}\|^2 \quad (47)$$

where $C_9 = max\{2C_8C_8^2\}$. A combining this with (46), we have

$$\sum_{k=1}^{N}\left\|\sum_{t=1}^{J}p_k^{n,t}\right\|^2 \geq \gamma\eta_n\sum_{k=1}^{N}\sum_{t=1}^{J}\|p_k^{n,t}\|^2$$
$$\geq \frac{\gamma\eta_n}{1 + C_9\eta_n(1+\eta_n)}\sum_{k=1}^{N}\sum_{t=1}^{J}\|p_k^{n,t}\|^2 \quad (48)$$

On the other hand, it follows from (43) and $0 \leq \eta_n \leq \eta_0 \leq 1$ that

$$\sum_{k=1}^{N}\left(\sum_{t=1}^{J}\|p_k^{n,t}\|\right)^2 \leq J\sum_{k=1}^{N}\sum_{t=1}^{J}\|p_k^{n,t}\|^2$$
$$\leq \frac{1}{\gamma\eta_n}\sum_{k=1}^{N}\left\|\sum_{t=1}^{J}p_k^{n,t}\right\|^2 \quad (49)$$

and

$$\eta_n\sum_{k=1}^{N}\sum_{t=1}^{J}\|p_k^{n,t}\| \leq \sqrt{\frac{J\eta_n}{\gamma}}\sum_{k=1}^{N}\left\|\sum_{t=1}^{J}p_k^{n,t}\right\|$$
$$\leq \sqrt{\frac{J}{\gamma}}\sum_{k=1}^{N}\left\|\sum_{t=1}^{J}p_k^{n,t}\right\| \quad (50)$$

This together with (19) yields

$$\|d_i^{n,j}\| \leq \eta_n\sum_{k=1}^{N}\left\|\sum_{t=1}^{J}p_k^{n,t}\right\| + C_2\eta_n^2\sum_{k=1}^{N}\sum_{t=1}^{J}\|p_k^{n,t}\|$$
$$\leq \left(1 + C_2\sqrt{\frac{J}{\gamma}}\right)\eta_n\sum_{k=1}^{N}\left\|\sum_{t=1}^{J}p_k^{n,t}\right\| \quad (51)$$

By (48) and (38), we deduce that

$$\sum_{j=1}^{J}p_i^{n+1,j} = \sum_{j=1}^{J}p_i^{n,j} + \sum_{j=1}^{J}r_i^{n,j} \quad (52)$$

It follows from (25), (45) and (51) that

$$\left\|\sum_{j=1}^{J}p_i^{n+1,j}\right\| \geq \sum_{j=1}^{J}\|p_k^{n,j}\| - J\|r_i^{n,j}\|$$
$$\geq \left\|\sum_{j=1}^{J}p_i^{n,j}\right\| - JC_9\eta_n\sum_{j=1}^{J}\|p_i^{n,j}\|$$
$$\geq \left(1 - JC_8\left(1 + C_2\sqrt{\frac{J}{\gamma}}\right)\eta_n\right)\sum_{k=1}^{N}\left\|\sum_{t=1}^{J}p_k^{n,t}\right\| \quad (53)$$

It can be easily verified that for any positive $x \geq y - z$, then

$$x^2 \geq y^2 - 2yz \quad (54)$$

Substitution (54) into(53) and noting (48), there holds

$$\left\|\sum_{j=1}^{J}p_k^{n+1,j}\right\|^2 \geq (1 + C_9\eta_n)\sum_{k=1}^{N}\left\|\sum_{t=1}^{J}p_k^{n,t}\right\|^2$$
$$\geq \frac{\gamma\eta_n(1 + C_{10}\eta_n)}{1 + C_9\eta_n(1+\eta_n)}\sum_{k=1}^{N}\sum_{t=1}^{J}\|p_k^{n+1,t}\|^2 \quad (55)$$

where $C_{10} = 2JC_8\left(1 + C_2\sqrt{J/\gamma}\right)$. Comparing (47) with (58), we see that if

$$\frac{\gamma\eta_n(1 + C_{10}\eta_n)}{1 + C_8\eta_n(1+\eta_n)} \geq \gamma\eta_{n+1} \quad (56)$$

From this easy to get (44) is proved. Hence we need only to verify (56). Substituting (9) into (56), we get

$$\beta - C_8 - C_{10} \geq (C_8 + \beta C_{10})\eta_n \quad (57)$$

Recalling the definition of $\tilde{\beta} = C_8 + C_9$ in (11), we see that if η_0 and β in (9) satisfy the conditions in Assumption (A3)

$$\beta > max\{1, \tilde{\beta}\} \text{ and } 0 \leq \eta_0 \leq min\left\{1, \frac{1}{\tilde{\beta}} - \frac{1}{\beta}\right\} \quad (58)$$

There holds

$$0 \leq \eta_n \leq \eta_0 \leq \frac{1}{\tilde{\beta}} - \frac{1}{\beta} = \frac{\beta - \tilde{\beta}}{\beta\tilde{\beta}} \leq \frac{\beta - C_8 - C_{10}}{C_8 + \beta C_{10}} \quad (59)$$

Which validates (57) and also (56). Thus, the inequality (44) has been proved.

The next two Lemmas will be used to prove our convergence results. Their proofs are omitted since they are quite similar to those of Lemma 3.5 in [27] and Theorem 3.5.10 in [28] , respectively

Lemma 7. Suppose that the series $\sum_{n=1}^{\infty} a_n^2/n < \infty$, that $a_n > 0\ for\ n = 1,2,....$ and that there exists a constant $\mu > 0$ such that $|a_{n+1} - a_n| < \mu/n\ , n = 1,2,...$ then, we have $\lim_{n\to\infty} a_n = 0$.

Lemma 8. Let $F: \Phi \subset R^p \to R\ (p \geq 1)$ be continuous for a bounded closed region Φ. If the set $\Phi_0 = \{x \in \Phi$: Fxx=0 has finite points and the sequence $xn \in \Phi$ satisfy:

(i) $\lim_{n\to\infty}\|F_x(x_n)\| = 0$

(ii) $\lim_{n\to\infty}\|x_{n-1}-x_n\|=0$.

Then, there exists $x^*\in\Phi_0$ such that $\lim_{n\to\infty}x_n=x^*$

Now we are ready to prove the main Theorems.

Proof of Theorem 1. In virtue of (37), if for any nonnegative integer n

$$\sum_{k=1}^{N}\left\|\sum_{t=1}^{J}p_k^{n,t}\right\|^2\geq\gamma\eta_n\sum_{k=1}^{N}\sum_{t=1}^{J}\|p_k^{n,t}\|^2 \tag{60}$$

then Theorem 1 is proved.

For $n=0$, if the left hand side of (60) is zero, then by (2) $E_{w_i}(w^0)=\sum_{i=1}^{J}p_i^{n,i}=0$. Hence, we have already reached a local minimum of the error function, and the iteration can be terminated. Otherwise, if $\|\sum_{i=1}^{J}p_i^{n,i}\|\neq 0$ Such that

$$\sum_{k=1}^{N}\left\|\sum_{t=1}^{J}p_k^{0,t}\right\|^2\geq\gamma\eta_n\sum_{k=1}^{N}\sum_{t=1}^{J}\|p_k^{0,t}\|^2 \tag{61}$$

Recalling Lemma 8, we know that Inequality(60) holds for all nonnegative integer n. Hence, the monotonicity of the error sequence $\{E(W^{nJ})\}$ is proved.

Proof of Theorem 2. Note that $\{\xi^{n1},\xi^{n2},\dots,\xi^{nJ}\}$ is the permutation of $\{\xi^1,\xi^2,\dots,\xi^J\}$ in the *n-th* cycle of training iteration, there holds for any $w\in\mathbb{R}^p$ and $n=0,1,\dots$ that

$$\begin{aligned}E(W^{nJ})&=\sum_{j=1}^{J}g_{nj}(\psi_i^{nJ})+\frac{\lambda}{2}\sum_{j=1}^{J}\sum_{i=1}^{N}(w_i^{nJ}\cdot\xi^{nj})^2\\&=\sum_{j=1}^{J}g_j(\psi_i^{nJ})+\frac{\lambda}{2}\sum_{j=1}^{J}\sum_{i=1}^{N}(w_i^{nJ}\cdot\xi^{j})^2\end{aligned} \tag{62}$$

Form Lemma 6, write

$$\begin{aligned}E(W^{nJ})\leq E(W^0)&=\sum_{j=1}^{J}g_j(\psi_i^0)+\frac{\lambda}{2}\sum_{j=1}^{J}\sum_{i=1}^{N}(w_i^0\cdot\xi^j)^2\\&\leq C_{14}\end{aligned} \tag{63}$$

where $C_{11}=JC+(\lambda/2)JC^2\|w_i^0\|^2$. From (1) and (63), we have

$$\lambda(w_i^{nJ}\cdot\xi^j)^2\leq 2E(W^{nJ})\leq 2C_{11},j=1,2,\dots,J \tag{64}$$

This together with the definition of C_6 in (25) indicates

$$\sum_{k=1}^{N}\sum_{t=1}^{J}\|p_k^{n,t}\|\leq C_6+\frac{2}{\lambda}C_{11} \tag{65}$$

Combining (2) with (62) we have

$$\begin{aligned}w_i^{nJ}=w_i^0-\eta_n\sum_{m=1}^{n-1}\sum_{j=1}^{J}\Big(&g_j'(\psi_i^{mJ+j-1})(\varphi_i^{mJ+j-1})\\&+\lambda(w_i^{mJ+j-1}\cdot\xi^j)\Big)\xi^j\end{aligned} \tag{66}$$

Let the second part of above equation be w_{i1}^{nJ}, Denote $\mathbb{R}_1=span\{\xi^1,\xi^2,..,\xi^J\}\subset\mathbb{R}^n$ and $\mathbb{R}_2=\mathbb{R}_1^{\perp}$ be the orthogonal complement space of $\mathbb{R}_1$. Denote the second part of (66) by w_{i1}^{nJ}, obviously $w_{i1}^{nJ}\in\mathbb{R}_1$. we divide w_i^0 into $w_i^0=w_{i1}^0+w_{i2}^0$, where $w_{i1}^0\in\mathbb{R}_1$ and $w_{i2}^0\in\mathbb{R}_2$. Then $w_i^n=(w_{i1}^0+w_{i1}^{nJ})\oplus w_{i2}^0=\widetilde{w}_{i1}^{nJ}\oplus w_{i2}^0$. Applying this to (66), we have

$$|s_t|:=|\widetilde{w}_{i1}^{nJ}\cdot\xi^t|=|w_i^{nJ}\cdot\xi^j|\leq\sqrt{\frac{C_{11}}{\lambda}},t=1,2,..,T \tag{67}$$

Suppose $\{\xi^{j_1},\xi^{j_2},..,\xi^{j_T}\}$ $(j_t\in\{1,\dots,J\},t=1,2,\dots,T)$ is a base of the space $\mathbb{R}_1$. There are $a_t\in\mathbb{R}$ $(t=1,2,\dots,T)$ such that $\widetilde{w}_{i1}^{nJ}=a_1\xi^{j_1}+\cdots+a_T\xi^{j_T}$. Then $(a_1\xi^{j_1}+\cdots+aT\xi jT\cdot\xi jt=st,t=1,\dots,T.$ we get

$$\begin{pmatrix}\xi^{j_1}\cdot\xi^{j_1}&\dots&\xi^{j_t}\cdot\xi^{j_1}\\\vdots\ \ \vdots&\vdots&\vdots\ \ \vdots\\\xi^{j_1}\cdot\xi^{j_t}&\dots&\xi^{j_t}\cdot\xi^{j_t}\end{pmatrix}\begin{pmatrix}a_1\\\vdots\\a_T\end{pmatrix}=\begin{pmatrix}s_1\\\vdots\\s_T\end{pmatrix} \tag{68}$$

Is a base, the coefficient determinant equal to zero, and the system of the linear equations has a unique solution. Assume that the coefficient determinant equals to:

$$C=\begin{vmatrix}\xi^{j_1}\xi^{j_1}\dots\xi^{j_{t-1}}\cdot\xi^{j_1}&\boldsymbol{d_1}&\xi^{j_{t-1}}\cdot\xi^{j_1}&\dots&\xi^{j_1}\cdot\xi^{j_1}\\\vdots\ \ \vdots\ \ \vdots&\vdots&\vdots&\vdots&\vdots\ \ \vdots\\\xi^{j_1}\xi^{j_t}\dots\xi^{j_{t-1}}\cdot\xi^{j_t}&\boldsymbol{d_t}&\xi^{j_{t-1}}\cdot\xi^{j_t}&\dots&\xi^{j_1}\ \ \xi^{j_T}\end{vmatrix}$$

Then the solution is as follows

$$a_t=C\cdot S^{-1} \tag{69}$$

Let the maximum absolute value of all the sub determinant with rank $(T-1)$ of the coefficient determinant is S', then $|a_t|\leq|S'|\cdot|S^{-1}|\cdot\sum_{t=0}^{T}|s_t|$. By (67) we have $|a_t|\leq|S'|\cdot|S^{-1}|\cdot T\cdot\sqrt{2C_{11}/\lambda}$. $t=1,2,\dots,T$. Denote $C_{11}'=\max_{1\leq t\leq T}\|\xi^{j_1}\|$, then

$$\begin{aligned}\|\widetilde{w}_{i1}^{nJ}\|&=\|a_1\xi^{j_1}+\cdots+a_T\xi^{j_T}\|\\&\leq|S'|\cdot|S^{-1}|\cdot C_{11}'\cdot T^2\cdot\sqrt{\frac{2C_{11}}{\lambda}}\end{aligned} \tag{70}$$

That is $\widetilde{w}_{i1}^{nJ}$ are bounded uniformly bounded. So from (67), we know w_i^{nJ} are uniformly bounded. In all, we get $\{w_i^{nJ}\}_{n=0}^{\infty}$ are uniformly bounded, i.e., there exist a bounded closed region $S\subset\mathbb{R}^n$ such that $\{w_i^{nJ}\}\subset S$.

Proof of Theorem 3. Denote

$$\sigma^n=\eta_n\sum_{k=1}^{N}\left\|\sum_{t=1}^{J}p_k^{n,t}\right\|^2-\gamma\eta_n^2\sum_{k=1}^{N}\sum_{t=1}^{J}\|p_k^{n,t}\|^2 \tag{71}$$

We observe from the proof of Theorem 1 that $\sigma^n\geq 0$ for $\forall\, n=0,1,\dots$.

In view Lemma 7 and Theorem 1, there holds

$$E(W^{(n+1)J})\leq E(W^{nJ})-\sigma^n\leq\cdots\leq E(W^0)-\sum_{k=1}^{n}\sigma^k \tag{72}$$

Note that $E(W^{(n+1)J})\geq 0$ for any $n>0$. Setting $n\to\infty$, we have

$$\sum_{n=0}^{\infty}\sigma^n\leq E(W^{nJ})<\infty \tag{73}$$

A combination of (65) and Lemma $2(ii)$ gives

$$\sum_{n=0}^{\infty}\left(\gamma\eta_n^2\sum_{k=1}^{N}\sum_{t=1}^{J}\|p_k^{n,t}\|^2\right)\leq C_{12}\sum_{n=0}^{\infty}\eta_n^2$$

$$< \sigma^n \sum_{n=1}^{\infty} \frac{C_{12}}{\eta_n} < \infty, \tag{74}$$

where $C_{13} = \gamma J C_{12}^2$. Thus and (73) holds

$$\sum_{n=0}^{\infty} \eta_n \sum_{k=1}^{N} \left\| \sum_{t=1}^{J} p_k^{n,t} \right\|^2 < \infty \tag{75}$$

Thus

$$\sum_{n=1}^{\infty} \frac{1}{n} \| E(W^{nJ}) \|^2 < \frac{1}{\tau} \sum_{n=0}^{\infty} \eta_n \sum_{k=1}^{N} \left\| \sum_{t=1}^{J} p_k^{n,t} \right\|^2 < \infty \tag{76}$$

Let $E_{ww}(W) = \left\{ \partial^2 E / \partial_{w_i} \partial_{w_j} \right\}_{1 \le i,j \le p}$ be the Hessian matrix of $E(W)$. By (A1),theorem 2 and Lemma 2, there is $C_{13} > 0$ such that

$$\left\| E_{w_i w_i}(W) \right\| = \left\| E_{w_i w_i}(W) \right\| < C_{14}, \quad w \in \mathbb{R}^p \tag{77}$$

In addition, by (33) and (65) , there is $C_{15} > 0$ such that

$$\left\| d_i^{n,j} \right\| \le (1 + C_2)\ \eta_n \sum_{k=1}^{N} \sum_{t=1}^{J} \left\| p_k^{n,t} \right\| < \frac{C_{15}}{n} \tag{78}$$

Again using the Taylor Expansion, holds

$$\begin{aligned} &\left\| \left\| E_{w_i}\left(W^{(n+1)J}\right) \right\| - \left\| E_{w_i}(W^{nJ}) \right\| \right\| \\ &\le \left\| E_{w_i}\left(W^{(n+1)J}\right) - E_{w_i}(W^{nJ}) - E_{w_i w_i}(W^{nJ}) d_i^{n,j} \right\| \\ &\qquad + \left\| E_{w_i w_i}(W^{nJ}) d_i^{n,j} \right\| \\ &\le \left(0\left(\left\| d_i^{n,j} \right\| \right) + C_{14} \left\| d_i^{n,j} \right\| \right) < C_{16} \left\| d_i^{n,j} \right\| < \frac{C_{17}}{n} \end{aligned} \tag{79}$$

where $C_{17} = C_{14} C_{16}$. A combination of (78), (79) and Lemma 7 gives

$$\lim_{n \to \infty} \left| E_{w_i}(W^{nJ}) \right| = 0. \tag{80}$$

Similarly as (79), there $C_{18} > 0$ for any unit vector, gives

$$\left\| E_{w_i}(W^{nJ+j}) - E_{w_i}(W^{nJ}) \right\| < \frac{C_{18}}{n}, j = 1,2, \dots, J \tag{81}$$

Thus

$$\begin{aligned} \left\| E_{w_i}(W^{nJ+j}) \right\| &\le \left\| E_{w_i}(W^{nJ}) \right\| \\ &\quad + \left\| E_{w_i}(W^{nJ+j}) - E_{w_i}(W^{nJ}) \right\| \\ &< \left\| E_{w_i}(W^{nJ}) \right\| + \frac{C_{18}}{n} \to 0 \end{aligned} \tag{82}$$

Namely, we come to the weak convergence result:

$$\lim_{k \to \infty} \left\| E_{w_i}(W^k) \right\| = 0 \tag{83}$$

Next, we prove the strong convergence. By (80), we get

$$\lim_{m \to \infty} \left\| w_i^{nJ+j} - w_i^{nJ} \right\| = \lim_{n \to \infty} \left\| d_i^{n,j} \right\| = 0 \tag{84}$$

Recalling Lemma 8 and noting (80) , (84) and Assumption (A3) there exists $w^* \in \Omega_0$ such that

$$\lim_{m \to \infty} W^{nJ} = W^*, \quad \left\| E_{w_i}(w_i^*) \right\| = 0 \tag{85}$$

Note that for $j = 1,2, \dots, J$, there is $C_{19} > 0$ such that

$$\left\| w_i^{nJ+j} - w_i^{nJ} \right\| \le \sum_{j=1}^{J} \left\| \left(\alpha_i^{n,j}\ \Delta_j^n w_i^{nJ+j-1} - \eta_n p_i^{n,j,j} \right) \right\| \tag{86}$$

$$\le \sum_{k=1}^{N} \sum_{t=1}^{J} \left\| \alpha_k^{n,t}\ \Delta_t^n w_k^{nJ+t-1} - \eta_n p_k^{n,t,t} \right\| \le C_{19} \eta_n \to 0 \tag{87}$$

Combining this with (85) yields

$$\lim_{n \to \infty} \left\| w_i^{nJ+j} - w_i^* \right\| = 0, \quad j = 1,2, \dots J \tag{88}$$

Hence

$$\lim_{k \to \infty} W^k = W^*, \quad \left\| E_{w_i}(W^*) \right\| = 0 \tag{89}$$

which completes the proof.

V. Conclusion

In this paper, we study boundedness and convergence of batch gradient method with inner-penalty and momentum for Pi-sigma neural network. The penalty term it is celled inner-penalty and it is useful to prove capability and magnitude network training. The momentum of the error function with penalty term is often insert to the increment formula for the weights so that the new weight updating rule prove increment during the training iteration. In this way, the network tends to respond not only to the local gradient but also to recent trends in the error surface. Both weak and strong convergence of the algorithm are considered for the net- work with a weights on the connections between the product node and the summation nodes are fixed to 1, which is the fast process during the training iteration. sufficient conditions for this convergence results are offered. Under this condition, we prove that the error function is decreasing monotonically, and the batch gradient method with inner-penalty and momentum is deterministically convergent.

Acknowledgment

We gratefully acknowledge to thank the anonymous referees for their valuable comments and suggestions on the revision of this paper. Special thanks to Prof. Dr. Xiong Yan for their kind helps during the period of the research.

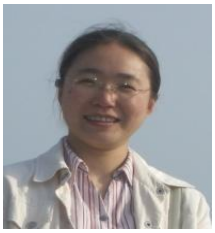
References

1. Y. Shin, J. Ghosh, The pi-sigma network: an efficient higher-order neural network for pattern classification and function approximation. International Joint Conference on Neural Networks, vol. 1(1991)13-18.
2. Y. Shin, J. Ghosh and Y. M. Yoon, A complex pi-sigma network and its application to equalization of nonlinear satellite channels, International Conference on Neural Networks vol. 1(1997)148-152.
3. X. Yu, M. O. Efe and O. Kaynak, A general back-propagation algorithm for feedforward neural networks learning, IEEE Trans. Neural Networks vol. 13(2002) 251-259.
4. P. Estevez and Y. Okabe, Training the piecewise linear-high order neural network through error back propagation, International Joint Conference on Neural Networks, vol. 1(1991) 711-716.
5. H. M. Shao and G. F. Zheng, Boundedness and convergence of online gradient method with penalty and momentum, Neurocomputing , vol. 74(2011) 765-770.
6. R. Reed, Pruning algorithms-a survey, IEEE Transactions on Neural Networks, vol. 8(1997) 185-204.
7. L. Villalobos and F. L. Merat, Learning capability assessment and feature space optimization for higher-order neural networks, IEEE Transaction on Neural Networks vol. 6 (1995) 267-272.
8. L. Ma and K. Khorasani, A new strategy for adaptively constructing multilayer feedforward neural networks, Neurocomputing vol. 51 (2003) 361-385.
9. A. J. Hussain and P. Liatsis, A new recurrent polynomial neural network for predictive image coding, Image Processing And Its Applications vol. 1 (1999) 82-86.
10. V. K. Asari, Training of a feedforward multiple-valued neural networks by error backpropagation with a multilevel threshold function, IEEE Trans. Neural Networks vol. 12 (2001) 1519-1521.
11. G. E.Hinton, Connectionist learning procedures, Artificial Intelligence, vol. 40 (1989) 185-234.

12. M.T. Hagan and M. B. Mehnaj, Training feedforward networks with Marquardt algorithms, IEEE Trans. Neural Networks vol. 5 (1994) 989-993.
13. S. Roy and J. J. Shynk, Analysis of the momentum LMS algorithm, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 38 (12) (1990) 2088-2098.
14. M. Torii and M.T. Hagan, Stability of steepest descent with momentum for quadratic functions, IEEE Transactions on Neural Networks, vol. 13 (3) (2002) 752-756.
15. J. Kong and W. Wu, Online gradient methods with a punishing term for neural networks. Northeast Math. J vol. 173 (2001) 371-378.
16. A. Crema, M. Loreto and M. Raydan, Spectral projected subgradient with a momentum term for the Lagrangean dual approach, Computers and Operations Research vol. 34 (10) (2007) 3174-3186.
17. V. V. Phansalkar and P. S. Sastry, Analysis of the back-propagation algorithm with momentum, IEEE Transactions on Neural Networks, vol. 5 (3) (1994) 505-506.
18. S. Abid, F. Fnaiech and M. Najim, A fast feedforwardt raining algorithm using a modified form of the standard back-propagation algorithm, IEEE Trans. Neural Networks, vol. 12 (2001) 424-430.
19. E. Istook, T. Martinez, E. Istook and T. Martinez, Improved backpropagation learning in neural networks with windowed momentum, International Journal of Neural System, vol. 12 (2002) 303-318.
20. R. Setiono, Apenalty-function approach for pruning feedforward neural networks, Neural Networks, vol. 9, (1997) 185-204.
21. A. Bhaya and E. Kaszkurewicz, Steepest descent with momentum for quadratic functions is a version of the conjugate gradient method, Neural Networks, vol. 17 (2004) 65-71.
22. H. M. Shao, D. P. Xu, G. F. Zheng and L. J. Liu, Convergence of an online gradient method with inner-penalty and adaptive momentum, Neurocomputing , vol. 77(2012) 243-252.
23. N.M. Zhang, W. Wu and G.F. Zheng, Convergence of gradient method with momentum for two-layer feedforward neural networks, IEEE Trans. Neural Networks vol. 17 (2) (2006) 522-525.
24. Z.G. Zeng, Analysis of Global Convergence and Learning Parameters of the Back-propagation Algorithm for Quadratic Functions, Lecture Notes in Computer Science, vol. 4682 (2007) 7-13.
25. N.M. Zhang, An online gradient method with momentum for two-layer feedforward neural networks, Appl. Math. Comput. Vol. 212(2009) 488-498.
26. N. M. Zhang, Deterministic Convergence of an Online Gradient Method with Momentum, Lecture Notesin Computer Science, vol.4113 (2006) 94-105.
27. W. Wu, G. R. Feng and X. Li, Training multilayer perceptrons via minimization of sum of ridge functions. Adv. Comput. Math., vol. 17 (2002) 331-347.
28. Y.X. Yuan and W.Y. Sun, Optimization Theory and Methods, Science Press, Beijing, 2001.

AUTHOR PROFILE

Kh. Sh. Mohamed received the M.S. degree from Jilin University, Changchun, China, in applied mathematics in 2011. He works as a lecturer of mathematics at College of Science Dalanj University, since 2011. Now he is working toward Ph.D. degree in computational mathematics at Dalian University of Technology, Dalian, China. His research interests include theoretical analysis and regularization methods for neural networks.



Yan Xiong received M.S. degree from Northeastern University, Shenyang, China, in 2003 and the Ph.D. from Dalian University of Technology, Dalian, China, in 2007. Her research interests include the convergence analysis and structural optimization of neural networks, especially in higher order neural networks.

Zhengxue Li received the Ph.D. degree in mathematics from Jilin University, Changchun, China, in 2001. He is currently an associate professor with Dalian University of Technology. His current research interests include nonlinear algorithm analysis and intelligent informa- tion processing.

Habtamu Z. A received the bachelor degree from Wollega University, Ethiopia, in mathematics education in 2009 and his Master's degree in Mathematics education from Addis Ababa University, Ethiopia in 2011. He worked as a lecturer of applied mathemat-ics at Assosa University, Ethiopia. Currently heis working toward Ph.D. degree in Applied mathematics at Dalian University of Technology, Dalian, China. His research interests include numerical optimization methods and neural networks.

Abdrhaman. M. Adam, received the M.Sc in Industrial and Computational Mathematics from University of Khart- oum, Khartoum –Sudan in 2009. He was also awarded B.Sc degree in Mathe- matics from Omdurman Islamic University, Omdurman-Sudan in 2003. He has works as a Lecturer of Mathematics at Faculty of Science and Technology, Omdurman Islamic University, Omdurman- Sudan since 2004. Currently, he is a PhD student in Computational Mathematics at Dalian University of Technology, Dalian – China, since 2012. His research interest lies in theoretical approach for the analysis and improvement of learning algorithms in solving nonlinear differential equations.