English Series NUMERICAL MATHEMATICS Vol.14, No.1
A Journal of Chinese Universities Feb. 2005

CONVERGENCE OF ONLINE GRADIENT METHOD WITH A PENALTY TERM FOR FEEDFORWARD NEURAL NETWORKS WITH STOCHASTIC INPUTS[*]

Shao Hongmei(邵红梅) Wu Wei(吴 微) Li Feng(李 峰)

Abstract *Online gradient algorithm has been widely used as a learning algorithm for feedforward neural network training. In this paper, we prove a weak convergence theorem of an online gradient algorithm with a penalty term, assuming that the training examples are input in a stochastic way. The monotonicity of the error function in the iteration and the boundedness of the weight are both guaranteed. We also present a numerical experiment to support our results.*

Key words *Feedforward neural network, Online gradient algorithm, Penalty term, Stochastic input, Convergence, Monotonicity, Boundedness.*

AMS(2000)subject classifications 68T15

1 Introduction

Online gradient algorithm (OGM) is commonly used for feedforward neural network (FNN) training [2,3,5,6]. The training is usually done by iteratively updating of the weights according to the error signal, which is the negative gradient of a sum-square error function (SSE). However, by using SSE as the error function sometimes the weight of the network becomes very large and the generalization performance is poor, even though the network is trained until the error on the training set is minimized. In order to resolve this problem, a popular choice is to add a penalty term to the standard error function [1,4,8,9]. When the training samples are trained

[*] Partly supported by the National Natural Science Foundation of China,and the Basic Research Program of the Committee of Science,Technology and Industry of National Defense of China.
Received: Sep. 15, 2004

in a fixed order, the effect of the penalty term in controlling the magnitude of the weight is investigated in [8,10]. However, the usual OGM chooses input ξ^i from the training samples $\{\xi^i, O^i\}$ in a stochastic order, which is important to help the training procedure to jump off from local minima. In this paper we shall show that, when input sample ξ^i is chosen in a specially stochastic order (cf.[7]), such an online gradient algorithm with a penalty term and stochastic inputs (POGM-S) is weakly convergent. Besides, the monotonicity of the error function in the training iteration and the boundedness of the weight are both guaranteed. We also present a simulation example to illustrate our results established in the paper. Experimental results indicate that, as well as being beneficial from controlling the magnitude of the weight, POGM-S makes the generalization performance of the network greatly improved.

For simplicity, a two-layer FNN is considered with N input nodes and one single output node. Assume that the transfer function $\sigma : \mathbb{R} \to \mathbb{R}$ is a pre-chosen sigmoid function, and denote the weight by $\omega = (w_1, \cdots, w_N)^T$. Suppose $\{\xi^i, O^i\}_{i=1}^J$ is the given set of training examples. Our error function with a penalty term has the form (cf.[8])

$$E(\omega) = \frac{1}{2}\sum_{i=1}^{J}\left(O^i - \sigma(\omega\cdot\xi^i)\right)^2 + \frac{\lambda}{2}\sum_{i=1}^{J}(\omega\cdot\xi^i)^2 \equiv \sum_{i=1}^{J}[f_i(\omega\cdot\xi^i) + \frac{\lambda}{2}(\omega\cdot\xi^i)^2], \tag{1.1}$$

where $\lambda > 0$ is the coefficient of the penalty term. Then the gradient function is given by

$$\nabla E(\omega) = \sum_{i=1}^{J}[f_i^{'}(\omega\cdot\xi^i) + \lambda(\omega\cdot\xi^i)]\xi^i. \tag{1.2}$$

Now we introduce the POGM-S algorithm. Let $\{\xi^{n1}, \xi^{n2}, \cdots, \xi^{nJ}\}$ be a stochastic permutation of $\{\xi^1, \xi^2, \cdots, \xi^J\}$ in the n-th cycle of the training iteration. Starting from an initial value ω^0, we proceed to refine it iteratively by the following rule

$$\omega^{nJ+i} = \omega^{nJ+i-1} + \triangle_i^n\omega^{nJ+i-1}, \quad i = 1, 2, \cdots, J; n = 0, 1, \cdots, \tag{1.3a}$$

$$\triangle_i^n\omega^{nJ+i-1} = -\eta_n[f_{ni}^{'}(\omega^{nJ+i-1}\cdot\xi^{ni}) + \lambda(\omega^{nJ+i-1}\cdot\xi^{ni})]\xi^{ni}, \tag{1.3b}$$

where η_n is the learning rate in the n-th training cycle. For an initial value $\eta_0 > 0$, η_n changes after each cycle of training iteration according to

$$\frac{1}{\eta_n} = \frac{1}{\eta_{n-1}} + \beta, \qquad n = 1, 2, \cdots, \tag{1.4}$$

where $\beta > 0$ is a constant. The following assumption is imposed throughout the paper.

Assumption 1 There is $C > 0$ such that for any $t \in \mathbb{R}$ and $1 \le i \le J$

$$|f_i(t)| \le C, \quad |f_i^{'}(t)| \le C, \quad |f_i^{''}(t)| \le C.$$

The rest of this paper is organized as follows. In Section 2 we present several preliminary lemmas. A monotonicity theorem, a boundedness theorem and a convergence theorem are established in Section 3. We also present an effective experiment in Section 4. We use $\|\cdot\|$ to stand for the Euclidean norm over $\mathbb{R}^N$, and C and C_i for genetic constants which may be different in different places.

2 Preliminary lemmas

For simplicity, we denote

$$r_{i,n} := \triangle_i^n \omega^{nJ+i-1} - \triangle_i^n \omega^{nJ}, \qquad i = 1, 2, \cdots, J; n = 0, 1, 2, \cdots, \tag{2.1}$$

$$\omega_d^n := \omega^{(n+1)J} - \omega^{nJ}, \qquad n = 0, 1, 2, \cdots. \tag{2.2}$$

Proofs of the following two lemmas can be found in [7].

Lemma 1 Let $\{\eta_n\}$be given by (1.4), there hold the following estimates for any $n = 1, 2, \cdots$

(1) $\eta_{n-1} > \eta_n > 0$;

(2) $\eta_n < \dfrac{\rho}{n}, \quad \rho = \dfrac{1}{\beta}$;

(3) $\eta_n > \dfrac{\tau}{n}$, where $\tau > 0$ is some constant;

(4) $\dfrac{\eta_{n+1}}{\eta_n} > \dfrac{1}{2}$.

Lemma 2 Suppose that the series $\displaystyle\sum_{n=1}^{\infty} \frac{a_n^2}{n} < \infty$, that $a_n > 0$ for $n = 1, 2, \cdots$, and that there exists a constant $\mu > 0$ satisfying $|a_{n+1} - a_n| < \dfrac{\mu}{n}$, $n = 1, 2, \cdots$, then we have $\lim\limits_{n\to\infty} a_n = 0$.

Lemma 3 If Assumption 1 holds, there is $C > 0$ such that

(1) $\displaystyle\omega^{nJ+i} = \omega^{nJ} + \sum_{k=1}^{i} (\triangle_k^n \omega^{nJ} + r_{k,n}), \qquad i = 1, 2, \cdots, J$

(2) $\displaystyle\sum_{i=1}^{J} \|r_{i,n}\| \le C\eta_n \sum_{i=1}^{J} \|\triangle_i^n \omega^{nJ}\|$

(3) $\displaystyle\|\omega_d^n\| \le C \sum_{i=1}^{J} \|\triangle_i^n \omega^{nJ}\|$

Proof Equation (1) can be directly derived from (1.3a) and (2.1). Particularly

$$\omega_d^n = \sum_{i=1}^{J} (\triangle_i^n \omega^{nJ} + r_{i,n}). \tag{2.3}$$

A combination of (2.10 (1.3b), Assumption 1 and the mean value theorem gives

$$\begin{aligned} \|r_{i,n}\| &\leq \eta_n|f_{ni}^{'}(\omega^{nJ+i-1}\cdot\xi^{ni})-f_{ni}^{'}(\omega^{nJ}\cdot\xi^{ni})|\|\xi^{ni}\|+\eta_n\lambda\|\omega^{nJ+i-1}-\omega^{nJ}\|\|\xi^{ni}\|^2 \\ &\leq C_1\eta_n\|f_{ni}^{''}(t_{i,n})(\omega^{nJ+i-1}-\omega^{nJ})\cdot\xi^{ni}\|+C_1\eta_n\|\omega^{nJ+i-1}-\omega^{nJ}\| \\ &\leq C_2\eta_n\|\omega^{nJ+i-1}-\omega^{nJ}\|, \end{aligned} \tag{2.4}$$

where $t_{i,n}\in\mathbb{R}^N$ lies on the segment between $\omega^{nJ+i-1}\cdot\xi^{ni}$ and $\omega^{nJ}\cdot\xi^{ni}$. Using Lemma 3(1) and Lemma 1(1) and proving by induction on $\|r_{k,n}\|$, we have

$$\|r_{i,n}\|\leq C_2\eta_n\left(\sum_{k=1}^{i-1}\|\triangle_k^n\omega^{nJ}\|+\sum_{k=1}^{i-1}\|r_{k,n}\|\right)\leq C_3\eta_n\sum_{k=1}^{i-1}\|\triangle_k^n\omega^{nJ}\|. \tag{2.5}$$

This leads to (2)

$$\sum_{i=1}^{J}\|r_{i,n}\|\leq C_3\eta_n\sum_{i=1}^{J}\sum_{k=1}^{i-1}\|\triangle_k^n\omega^{nJ}\|\leq C\eta_n\sum_{i=1}^{J}\|\triangle_i^n\omega^{nJ}\|. \tag{2.6}$$

It follows from (2.3) (2.6) and Lemma 1(1) that

$$\|\omega_d^n\|\leq\sum_{i=1}^{J}\|\triangle_i^n\omega^{nJ}\|+\sum_{i=1}^{J}\|r_{i,n}\|\leq C_1\sum_{i=1}^{J}\|\triangle_i^n\omega^{nJ}\|. \tag{2.7}$$

Applying Cauchy-Schwartz Inequality leads to (3), we have

$$\|\omega_d^n\|^2\leq\left(C_1\sum_{i=1}^{J}\|\triangle_i^n\omega^{nJ}\|\right)^2\leq C\sum_{i=1}^{J}\|\triangle_i^n\omega^{nJ}\|^2. \tag{2.8}$$

Lemma 4 Let Assumption 1 be satisfied and the sequence $\{\omega^{nJ+k}\}$ be generated by the algorithm (1.3), then there is a positive constant γ independent of n such that

$$E(\omega^{(n+1)J})\leq E(\omega^{nJ})-\frac{1}{\eta_n}\Big\|\sum_{i=1}^{J}\triangle_i^n\omega^{nJ}\Big\|^2+\gamma\sum_{i=1}^{J}\big\|\triangle_i^n\omega^{nJ}\big\|^2.$$

Proof Let $\{\xi^{n1},\xi^{n2},\cdots,\xi^{nJ}\}$ be a permutation of $\{\xi^1,\xi^2,\cdots,\xi^J\}$ in the n-th cycle of training iteration. Let $\xi^{(n+1)i}=\xi^{nk_i}$ $(1\leq i\leq J)$, where $\{k_1,k_2,\cdots,k_J\}$ is a stochastic permutation of the subscription index set $\{1,2,\cdots,J\}$. From (1.3) we see that

$$\sum_{i=1}^{J}\triangle_{k_i}^n\omega^{nJ}=\sum_{i=1}^{J}\triangle_i^n\omega^{nJ}. \tag{2.9}$$

Thus, using Taylor expansion and (1.1) (1.3b) (2.2) (2.3) (2.9) we obtain

$$\begin{aligned}
E(\omega^{(n+1)J}) &= \sum_{i=1}^{J}\big[f_{(n+1)i}(\omega^{(n+1)J}\cdot\xi^{(n+1)i})+\frac{\lambda}{2}(\omega^{(n+1)J}\cdot\xi^{(n+1)i})^2\big]\\
&= \sum_{i=1}^{J}\big[f_{nk_i}(\omega^{(n+1)J}\cdot\xi^{nk_i})+\frac{\lambda}{2}(\omega^{(n+1)J}\cdot\xi^{nk_i})^2\big]\\
&= \sum_{i=1}^{J}\big[f_{nk_i}(\omega^{nJ}\cdot\xi^{nk_i})+\frac{\lambda}{2}(\omega^{nJ}\cdot\xi^{nk_i})^2\big]+\sum_{i=1}^{J}\big[f'_{nk_i}(\omega^{nJ}\cdot\xi^{nk_i})(\omega_d^n\cdot\xi^{nk_i})\\
&\quad +\lambda(\omega^{nJ}\cdot\xi^{nk_i})(\omega_d^n\cdot\xi^{nk_i})\big]+\frac{1}{2}\sum_{i=1}^{J}[f''_{nk_i}(\tilde{t}_{n,k_i})(\omega_d^n\cdot\xi^{nk_i})^2+\lambda(\omega_d^n\cdot\xi^{nk_i})^2]\\
&= E(\omega^{nJ})-\frac{1}{\eta_n}\sum_{i=1}^{J}\triangle_{k_i}^n\omega^{nJ}\cdot\omega_d^n+\sum_{i=1}^{J}\delta_{k_i,n}^n\\
&= E(\omega^{nJ})-\frac{1}{\eta_n}\big\|\sum_{i=1}^{J}\triangle_i^n\omega^{nJ}\big\|^2+\delta_n^n,
\end{aligned} \tag{2.10}$$

where $\tilde{t}_{n,k_i}\in\mathbb{R}$ is a vector between $\omega^{(n+1)J}\cdot\xi^{nk_i}$ and $\omega^{nJ}\cdot\xi^{nk_i}$, $\quad i=1,2,\cdots,J$, and

$$\delta_{k_i,n}^n=\frac{1}{2}[f''_{nk_i}(\tilde{t}_{n,k_i})(\omega_d^n\cdot\xi^{nk_i})^2+\lambda(\omega_d^n\cdot\xi^{nk_i})^2],$$

$$\delta_n^n=-\frac{1}{\eta_n}(\sum_{i=1}^{J}\triangle_i^n\omega^{nJ})\cdot(\sum_{i=1}^{J}r_{i,n})+\sum_{i=1}^{J}\delta_{k_i,n}^n.$$

It follows from Assumption 1 and (2.8) that

$$|\delta_{k_i,n}^n|\leq C_1\|\omega_d^n\|^2<C\sum_{i=1}^{J}\|\Delta_i^n\omega^{nJ}\|^2. \tag{2.11}$$

A combination of (2.6) (2.11) and Cauchy-Schwartz Inequality produces

$$|\delta_n^n|\leq\frac{1}{\eta_n}\sum_{i=1}^{J}\|\triangle_i^n\omega^{nJ}\|\cdot\sum_{i=1}^{J}\|r_{i,n}\|+\sum_{i=1}^{J}\|\delta_{k_i,n}^n\|\leq\gamma\sum_{i=1}^{J}\|\triangle_i^n\omega^{nJ}\|^2. \tag{2.12}$$

(2.10) together with (2.12) gives

$$E(\omega^{(n+1)J})\leq E(\omega^{nJ})-\frac{1}{\eta_n}\big\|\sum_{i=1}^{J}\triangle_i^n\omega^{nJ}\big\|^2+\gamma\sum_{i=1}^{J}\big\|\triangle_i^n\omega^{nJ}\big\|^2. \tag{2.13}$$

3 Main results

Now we first present the monotonicity theorem, of which the proof is similar to that in [7] and omitted.

Theorem 5 (Monotonicity theorem) Let the error function $E(\omega)$ be given by (1.1) and Assumption 1 be valid. For any initial value $\omega^0\in\mathbb{R}^N$, if the initial value η_0 is chosen to satisfy

$$\frac{1}{\eta_0}\big\|\sum_{i=1}^{J}\triangle_i^0\omega^0\big\|^2\geq\gamma\sum_{i=1}^{J}\big\|\triangle_i^0\omega^0\big\|^2$$

then, the sequence $\{E(\omega^{kJ})\}$ generated from the algorithm (1.3) decreases monotonically, namely

$$E(\omega^{(n+1)J}) \leq E(\omega^{nJ}).$$

The next theorem confirms the boundedness of the weights in the training procedure.

Theorem 6 (Boundedness theorem) Under the same assumption of Theorem 5, the weight sequence $\{\omega^k\}$ generated by (1.3) is uniformly bounded.

Proof Note that $\{\xi^{n1}, \xi^{n2}, \cdots, \xi^{nJ}\}$ is the permutation of $\{\xi^1, \xi^2, \cdots, \xi^J\}$ in the n-th cycle of training iteration, there holds for any $\omega \in \mathbb{R}$

$$\sum_{i=1}^{J}[f_{ni}(\omega\cdot\xi^{ni}) + \frac{\lambda}{2}(\omega\cdot\xi^{ni})^2] = \sum_{i=1}^{J}[f_i(\omega\cdot\xi^i) + \frac{\lambda}{2}(\omega\cdot\xi^i)^2], \quad n = 0, 1, 2, \cdots. \tag{3.1}$$

According to Theorem 5, Assumption 1, and (1.1)(3.1) we have

$$E(\omega^{nJ}) \leq E(\omega^0) = \sum_{i=1}^{J}[f_i(\omega^0\cdot\xi^i) + \frac{\lambda}{2}(\omega^0\cdot\xi^i)^2] \leq M, \tag{3.2}$$

where

$$M = \sum_{i=1}^{J}\left(\sup_{1\leq i\leq J} f_i(\omega^0\cdot\xi^i) + \frac{\lambda}{2}\|\omega^0\|^2\|\xi^i\|^2\right).$$

From (1.1) (3.2) we get

$$\lambda(\omega^{nJ}\cdot\xi^i)^2 \leq 2E(\omega^{nJ}) \leq 2M, \quad i = 1, 2, \ldots, J. \tag{3.3}$$

Combining (1.3) with (3.1) we have

$$\omega^{nJ} = \omega^0 + \sum_{k=0}^{n-1}\sum_{i=1}^{J}\left\{-\eta_k[f_i^{'}(\omega^{kJ+i-1}\cdot\xi^i) + \lambda(\omega^{kJ+i-1}\cdot\xi^i)]\xi^i\right\}. \tag{3.4}$$

Let $A_1 = \text{span}\{\xi^1, \xi^2, \cdots, \xi^J\} \subset \mathbb{R}^n$ and $A_2 = A_1^{\perp}$ be the orthogonal complement space of A_1. Denote the second part of (3.4) by ω_1^{nJ}, obviously $\omega_1^{nJ} \in A_1$. We divide ω^0 into $\omega^0 = \omega_1^0 + \omega_2^0$, where $\omega_1^0 \in A_1$ and $\omega_2^0 \in A_1^{\perp}$. Then $\omega^{nJ} = (\omega_1^0 + \omega_i^{nJ}) \bigoplus \omega_2^0 \equiv \tilde{\omega}_1^{nJ} \bigoplus \omega_2^0$. Applying this to (3.3) we have

$$|d_k| := |\tilde{\omega}_1^{nJ}\cdot\xi^k| = |\omega^{nJ}\cdot\xi^k| \leq \sqrt{\frac{2M}{\lambda}}, \qquad k = 1, \ldots, K. \tag{3.5}$$

Suppose $\{\xi^{i_1}, \xi^{i_2}, \ldots, \xi^{i_K}\}$ $(i_k \in \{1, 2, \cdots, J\}, k = 1, 2, \cdots, K)$ is a base of the space A_1. There are $a_k \in \mathbb{R}(k = 1, 2, \cdots, K)$ such that $\tilde{\omega}_1^{nJ} = a_1\xi^{i_1} + \cdots + a_K\xi^{i_K}$. Then $(a_1\xi^{i_1} + \cdots + a_K\xi^{i_K})\cdot\xi^{i_k} = d_k$, $k = 1, 2, \cdots, K$. The matrix form is

$$\begin{pmatrix} \xi^{i_1}\cdot\xi^{i_1} & \cdots & \xi^{i_K}\cdot\xi^{i_1} \\ \vdots & \vdots & \vdots \\ \xi^{i_1}\cdot\xi^{i_K} & \cdots & \xi^{i_K}\cdot\xi^{i_K} \end{pmatrix}\begin{pmatrix} a_1 \\ \vdots \\ a_K \end{pmatrix} = \begin{pmatrix} d_1 \\ \vdots \\ d_K \end{pmatrix}. \tag{3.6}$$

Because $\{\xi^{i_1},\dots,\xi^{i_K}\}$ is a base, the coefficient determinant is not equal to zero, and the system of the linear equations has a unique solution. Assume that the coefficient determinant equals to D, then the solution is as follows

$$a_k = \begin{vmatrix} \xi^{i_1}\cdot\xi^{i_1} & \dots & \xi^{i_{k-1}}\cdot\xi^{i_1} & d_0 & \xi^{i_{k+1}}\cdot\xi^{i_1} & \dots & \xi^{i_K}\cdot\xi^{i_1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \xi^{i_1}\cdot\xi^{i_K} & \dots & \xi^{i_{k-1}}\cdot\xi^{i_K} & d_K & \xi^{i_{k+1}}\cdot\xi^{i_K} & \dots & \xi^{i_K}\cdot\xi^{i_K} \end{vmatrix}\cdot D^{-1}.$$

Let the maximum absolute value of all the subdeterminant with rank (K-1) of the coefficient determinant is D', then $|a_k| \le |D'|\cdot|D^{-1}|\cdot\sum\limits_{k=0}^{K}|d_k|$. By (3.5) we have $|a_k| \le |D'|\cdot|D^{-1}|\cdot K\cdot\sqrt{\dfrac{2M}{\lambda}},\quad k=1,2,\dots,K$. Denote $M' = \max\limits_{1\le k\le K}\|\xi^{i_k}\|$, then

$$\|\tilde\omega_1^{nJ}\| = \|a_1\xi^{i_1}+\dots+a_K\xi^{i_K}\| \le |D'|\cdot|D^{-1}|\cdot M'\cdot K^2\cdot\sqrt{\frac{2M}{\lambda}}. \tag{3.7}$$

So $\omega^{nJ} = \tilde\omega_1^{nJ}\bigoplus\omega_2^0$ is also uniformly bounded. Similarly, we can show for $i=1,2,\dots,J-1$ that $\{\omega^{nJ+i}\}_{n=0}^{\infty}$ are uniformly bounded. In all, the weight sequence $\{\omega^k\}_{k=0}^{\infty}$ is uniformly bounded.

Theorem 7 (Weak convergence theorem) Under the same assumptions of Theorem 5, we have

$$\lim_{k\to\infty}\|\nabla E(\omega^k)\| = 0.$$

Proof According to (2.13) we obtain

$$E(\omega^{(n+1)J}) \le \dots \le E(\omega^J) - \sum_{k=1}^{n}\left(\frac{1}{\eta_k}\|\sum_{i=1}^{J}\triangle_i^k\omega_{kJ}\|^2 - \gamma\sum_{i=1}^{J}\|\triangle_i^k\omega_{kJ}\|^2\right). \tag{3.8}$$

Since $E(\omega^{(n+1)J})\ge 0$, let $n\to\infty$ we get

$$\sum_{n=1}^{\infty}\left(\frac{1}{\eta_n}\|\sum_{i=1}^{J}\triangle_i^n\omega^{nJ}\|^2 - \gamma\sum_{i=1}^{J}\|\triangle_i^n\omega^{nJ}\|^2\right) \le E(\omega^J) < \infty. \tag{3.9}$$

A combination of (1.3), Assumption 1, Theorem 5 and Lemma 1 (2) gives

$$\|\triangle_i^n\omega^{nJ}\| \le \eta_n\big|f'_{ni}(\omega^{nJ}\cdot\xi^{ni}) + \lambda(\omega^{nJ}\cdot\xi^{ni})\big|\|\xi^{ni}\| \le C_1\eta_n < \frac{\rho C_1}{n}. \tag{3.10}$$

Thus

$$\sum_{n=1}^{\infty}\left(\gamma\sum_{i=1}^{J}\|\triangle_i^n\omega^{nJ}\|^2\right) \le \gamma\rho^2 JC_1^2\sum_{n=1}^{\infty}\frac{1}{n^2} < \infty. \tag{3.11}$$

It follows from (1.2) (1.3) (3.9) (3.11) and Lemma 1(3) that

$$\sum_{n=1}^{\infty}\frac{1}{n}\big\|\nabla E(\omega^{nJ})\big\|^2 < \frac{1}{\tau}\sum_{n=1}^{\infty}\frac{1}{\eta_n}\big\|\sum_{i=1}^{J}\triangle_i^n\omega^{nJ}\big\|^2 < \infty. \tag{3.12}$$

Let $\nabla^2 E(\omega) = \left\{ \dfrac{\partial^2 E}{\partial\omega_i \partial\omega_j} \right\}_{1\leq i,j\leq N}$ be the Hessian matrix of $E(\omega)$. Then by (2.7) (3.10) and Assumption 1, there is $C_1 > 0$ such that

$$\|\nabla^2 E(\omega)\| < C_1, \qquad \|\omega_d^n\| < \frac{C_1}{n}. \tag{3.13}$$

Again using Taylor expansion and noting (3.13) we have

$$\begin{aligned} &\big| \|\nabla E(\omega^{(n+1)J})\| - \|\nabla E(\omega^{nJ})\| \big| \\ \leq& \big\| \nabla E(\omega^{(n+1)J}) - \nabla E(\omega^{nJ}) \big\| \\ \leq& \big\| \nabla E(\omega^{(n+1)J}) - \nabla E(\omega^{nJ}) - \nabla^2 E(\omega^{nJ})\omega_d^n \big\| + \big\| \nabla^2 E(\omega^{nJ})\omega_d^n \big\| \\ \leq& \big(o(\|\omega_d^n\|) + C_1\|\omega_d^n\| \big) < C_2\|\omega_d^n\| < \tfrac{C}{n}. \end{aligned} \tag{3.14}$$

By (3.12) (3.14) and Lemma 2, we conclude

$$\lim_{n\to\infty} \|\nabla E(\omega^{nJ})\| = 0. \tag{3.15}$$

Similarly as (3.14), we have

$$\|\nabla E(\omega^{nJ+i}) - \nabla E(\omega^{nJ})\| < \frac{C}{n}, \qquad i = 1, 2, \cdots, J. \tag{3.16}$$

Thus, (3.15) together with (3.16) gives

$$\begin{aligned} \|\nabla E(\omega^{nJ+i})\| \quad &\leq \quad \|\nabla E(\omega^{nJ})\| + \|\nabla E(\omega^{nJ+i}) - \nabla E(\omega^{nJ})\| \\ &< \quad \|\nabla E(\omega^{nJ})\| + \tfrac{C}{n} \to 0 \quad (n \to \infty), \quad i = 1, 2, \cdots, J. \end{aligned} \tag{3.17}$$

A combination of (3.17) (3.15) and Lemma 2 leads to the conclusion:

$$\lim_{k\to\infty} \|\nabla E(\omega^k)\| = 0,$$

which completes the proof.

Fig. 1 Square error and norm of gradient with penalty term

Fig.2 Norm of weight compared with no penalty term

4 Numerical experiments

To illustrate the capacity of the learning algorithm used in this paper, a pattern classification problem is considered. The training examples are

$$\{\xi^1 = (-1, 1), O^1 = 1\} \quad \{\xi^2 = (1, -1), O^2 = 0\}$$

$$\{\xi^3 = (-3, 2), O^3 = 1\} \quad \{\xi^4 = (2, -3), O^4 = 0\}$$

From Fig. 1, we can see that the square error decreases monotonically and the corresponding gradient tends to zero. The effectiveness of the algorithm in controlling the weight is shown in Fig. 2. Without the penalty term, the weight becomes larger and larger during the training iteration. After adding the penalty term, the magnitude of the weight becomes smaller and smaller, and finally tends to keep steady. Table.1 shows that the larger the coefficient λ is, the smaller the weight becomes. Hence, our approach provides a mechanism to effectively control the magnitude of the weights, which might be important for the neural networks.

Table.1 Effect of the coefficient λ on square error and weight

$\eta = 0.9$	Square error	$\|w\|$	$\eta = 0.5$	Square error	$\|w\|$
$\lambda = 0$	0.003096	5.894	$\lambda = 0$	0.005837	4.924
$\lambda = 0.001$	0.03249	2.487	$\lambda = 0.001$	0.03222	2.583
$\lambda = 0.002$	0.05718	1.785	$\lambda = 0.002$	0.05607	1.931
$\lambda = 0.003$	0.07897	1.389	$\lambda = 0.003$	0.07666	1.565
$\lambda = 0.004$	0.09887	1.12	$\lambda = 0.004$	0.09512	1.31
$\lambda = 0.005$	0.1174	0.9252	$\lambda = 0.005$	0.112	1.117
$\lambda = 0.006$	0.1347	0.7803	$\lambda = 0.006$	0.1276	0.9645
$\lambda = 0.007$	0.1511	0.6702	$\lambda = 0.007$	0.1422	0.8417
$\lambda = 0.008$	0.1666	0.5851	$\lambda = 0.008$	0.1559	0.7415
$\lambda = 0.009$	0.1814	0.5183	$\lambda = 0.009$	0.1688	0.6591
$\lambda = 0.01$	0.1955	0.4651	$\lambda = 0.01$	0.181	0.5907

References

1 Hinton G E. Connectionist learning procedures. Artificial Intelligence, 1989, 40: 185-234

2 Wu W, Feng G R, Li X. Training multilayer perceptrons via minimization of sum of ridge functions. Advances in Computational Mathematics, 2002, 17: 331-347

3 Gori M, Maggini M. Optimal convergence of online backpropagation. IEEE Trans. Neural Networks, 1996, 7: 251-254

4 Setiono R. A penalty-function approach for pruning feedforward neural networks. Neural Computation, 1997, 9: 185-204

5 Ellacott S W. The numerical approach of neural networks. Mathematical Approaches to Neural Networks, ed.J.G.Taylor (North-Holland, Amsterdam, 1993), 103-138

6 Hassoun M H. Foundamentals of Artificial Neural Networks. (MIT Press, Cambridge, MA, 1995)

7 Li Z X, Wu W, Tian Y L. Convergence of an online gradient method for feedforword neural networks with stochastic inputs. J. Comput. Appl. Math., 2004, 163: 165-176

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

9 Weigend A S, Rumelhart D E, Huberman B A. Generalization by weight-elimination applied to currency exchange rate prediction. Proc. Intl. Joint Conf. on Neural Networks 1 (Seatle, 1991) 837-841

10 Zhang L Q, Wu W. Online gradient methods with a penalty term for neural networks with large training set. J. Nolinear Dynamics, accepted, 2004

Shao Hongmei Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, PRC.

Wu Wei Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, PRC.

Li Feng Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, PRC.