Article ID: 1000-341X(2006)03-0451-06 **Document code**: A

Finite Convergence of On-line BP Neural Networks with Linearly Separable Training Patterns

SHAO Zhi-qiong, WU Wei, YANG Jie
(Dept. of Appl. Math., Dalian University of Technology, Liaoning 116023, China)
(E-mail: shaozhiqiong@126.com)

Abstract: In this paper we prove a finite convergence of online BP algorithms for nonlinear feedforward neural networks when the training patterns are linearly separable.

Key words: nonlinear feedforward neural networks; online BP algorithms; finite convergence; linearly separable training patterns.
MSC(2000): 68T01
CLC number: TP183

1. Introduction

One of the mathematical foundations of neural networks is the convergence of training algorithms. Some algorithms such as the perceptron rule[2] and the delta rule[3] have proved convergent for linearly separable training patterns. Researchers have also attempted to obtain the convergence of the online BP algorithm for nonlinear multilayer perceptrons. One of these attempts is Gori & Maggini[1] in which they try to prove a convergence result for online BP multilayer neural networks with linearly separable training patterns under certain assumptions. Unfortunately, their paper contains a mathematical mistake that renders the proofs erroneous [see the Appendix].

In [4] we have proved the convergence of online BP for single layer nonlinear perceptrons with linearly separable training patterns. The aim of this paper is to generalize the method in [4] to prove the convergence of online BP multilayer neural networks under some assumptions which are similar to those in [1], but stronger than those in [4]. This generalization is necessary since the BP multilayer neural network is used most-often in practice.

This paper is organized as follows. In the next section, we state some preliminaries. In Section 3, we present some lemmas and the convergence result. An appendix is attached in the end of the paper.

2. Preliminaries

Let the weight vectors of the networks be $W_i = (w_{i1}, \cdots, w_{im})^T \in R^m, i = 1, \cdots, p$, where w_{ij} denotes the weight connecting the jth input neuron and ith hidden neuron; and

Received date: 2004-10-10
Foundation item: the National Natural science Foundation of China (10471017), and the Basic Research Program of the National Defence Committee of Science, Technology and Industry of China (K1400060406)

$V = (v_1, \cdots, v_p)^T \in R^p$ where $v_i \in R^+$ (see Remark 1 below) is the weight between the ith hidden neuron and the output neuron. For an input vector $U = (u_1, \cdots, u_m)^T \in R^m$, the output of the ith hidden neuron is

$$x_i = f(a_i), a_i = \sum_{j=1}^{m} u_j w_{ij} = W_i^T U,$$

and the output of the network is

$$x_0 = f(a_0), a_0 = \sum_{k=1}^{p} v_k x_k = V^T H,$$

where $H = (x_1, \cdots, x_p)^T$ and $f(x) : R \to I$ $(I = (-1, 1))$ is a smooth sigmoidal function (for example $f(x) = \tan h(x)$). Such type of functions has the follow properties that will be employed in our future proofs.

Property 1 $\lim_{x\to\infty} f(x) = 1, \lim_{x\to-\infty} f(x) = -1$.

Property 2 $f(x)$ is an odd function: $f(-x) = -f(x)$.

Property 3 $\lim_{x\to\pm\infty} f'(x) = 0$.

Property 4 $\forall M > 0, \exists G_M > 0$, s.t. $f'(x) \geq G_M$ for $-M \leq x \leq M$.

The following properties are direct consequences of the above properties:

Property 5 $f'(x)$ is an even function: $f'(-x) = f'(x)$. (By property 2)

Property 6 $f(x)$ is strictly increasing, so the inverse function $f^{-1}(x)$ exists. (By property 4)

The neural network is supplied with a set of training pattern pairs $\{\xi^q, O^q\}_{q=1}^{\varrho} \subset R^m \times \{\pm 1\}$ which are arranged stochastically to form a sequence of input-target pairs $\{U^k, d^k\}_{k=0}^{\infty} \subset R^m \times \{\pm 1\}$, such that each pair $\{\xi^q, O^q\}$ appears infinite times.

Definition 1 *The set of training patterns is linearly separable, if there exist a vector $A \in R^m$ and a constant $C_1 > 0$ such that*

$$A^T \xi^q \begin{cases} \geq C_1, & \text{if } O^q = 1, \\ \leq -C_1, & \text{if } O^q = -1. \end{cases} \tag{1}$$

The error function is chosen as

$$E(W, V) = \frac{1}{2} \sum_{q=1}^{\varrho} (x_0^q - O^q)^2.$$

So for a given constant $\varepsilon > 0$, the weights are updated at the kth step of training as follows:

$$W_i^{k+1} = \begin{cases} W_i^k & \text{if } |d^k - x_0^k| < \varepsilon, \quad (2a) \\ W_i^k + \eta_i^k (d^k - x_0^k) f'(a_i^k) f'(a_0^k) v_i^k U^k, & \text{if } |d^k - x_0^k| \geq \varepsilon, \quad (2b) \end{cases}$$

$$V^{k+1} = \begin{cases} V^k, & \text{if } |d^k - x_0^k| < \varepsilon, \quad (3a) \\ V^k + \eta(d^k - x_0^k) f'(a_0^k) H^k, & \text{if } |d^k - x_0^k| \geq \varepsilon, \quad (3b) \end{cases}$$

where the learning rate $\eta_i^k = \mu / v_i^k$ is variable, and μ and η are positive constants.

Since what we are concerned is the actually refined weight vector W_i^k in (2b) and V^k in (3b), we can drop out those U^k and W_i^k that satisfy (2a) and V^k that satisfy (3a), and assume every W_i^k and V^k satisfies (2b) and (3b) respectively. If we set $\widetilde{U}^k = d^k U^k$, then $\{U^k, d^k\}_{k=0}^{\infty}$ corresponds to $\{\widetilde{U}^k, 1\}_{k=0}^{\infty}$. Let us still use U^k for $\widetilde{U}^k$. In these notations, the sequence of input-target pairs becomes $\{U^k, 1\}_{k=0}^{\infty}$ and according to (1) we have

$$A^T U^k \geq C_1, k = 0, 1, \cdots. \quad (4)$$

And now U^k, W_i^k and V^k satisfy

$$1 - x_0^k \geq \varepsilon, \quad (5)$$

$$W_i^{k+1} = W_i^k + \mu(1 - x_0^k) f'(a_i^k) f'(a_0^k) U^k, \quad (6)$$

$$V^{k+1} = V^k + \eta(1 - x_0^k) f'(a_0^k) H^k. \quad (7)$$

In fact, by Properties 2 and 5, we see that the weight sequences $\{W_i^k\}$ and $\{V^k\}$ remain unchanged under our simplification of symbols. In the sequel, we always assume (4)–(7).

For the training procedures (6) and (7), there are two cases to consider:

Case I. The training procedures (6) and (7) terminates in finite number of steps when the output x_0^q satisfies $|d^q - x_0^q| < \varepsilon$ for every training example ξ^q.

Case II. The training procedure (6) and (7) does not terminate in finite number of steps and we have two infinite sequences $\{W_i^k\}_{k=0}^{\infty}$ and $\{V_i^k\}_{k=0}^{\infty}$ satisfying (6) and (7).

We shall proceed by a contradiction argument in the sequel to show that we must have Case I to be valid. So until the last theorem we always assume Case II , or equivalently, assume the existence of the infinite sequences $\{a_0^k\}_{k=0}^{\infty}$ and $\{a_i^k\}_{k=0}^{\infty}$ satisfying (5)–(7).

3. Convergence of the on-line BP

The following two assumptions will be used in this paper (cf. Remarks 1 and 2):

(III). $v_i \in R^+$.

(IV). For some i_0, $\|W_{i_0}\| \leq C$ (C is a given constant).

So $a_{i_0}^k = W_{i_0}^T U^k$ is also bounded for any k. According to the assumption and Property 4 there exists a constant $C_2 > 0$, such that

$$f'(a_{i_0}^k) \geq C_2, \forall K = 1, 2, \cdots. \quad (8)$$

Lemma 1 *Assume Case II, then there exist a subsequence $\{a_0^{k_n}\}_{n=1}^{\infty}$ of $\{a_0^k\}_{k=0}^{\infty}$ in (7) and a constant M, such that $k_n \to \infty$ as $n \to \infty$, $a_0^k \geq M$ if $a_0^k \in \{a_0^{k_n}\}$, and $a_0^k < M$ if $a_0^k \notin \{a_0^{k_n}\}$.*

Proof Using (7), we have

$$\begin{aligned}\|V^{k+1}\|^2 &= \|V^k + \eta(1-x_0^k)f'(a_0^k)H^k\|^2 \\ &= \|V^k\|^2 + 2\eta(1-x_0^k)f'(a_0^k)a_0^k + \eta^2(1-x_0^k)^2 f'(a_0^k)^2\|H^k\|^2.\end{aligned}$$

Notice that $1-x_0^k, f'(a_0^k)$ and $\|H^k\|$ are positive and bounded for arbitrary k. Thus if $a_0^k < -M_1$, for a sufficiently large positive number M_1, there holds

$$2\eta(1-x_0^k)f'(a_0^k)a_0^k + \eta^2(1-x_0^k)^2 f'(a_0^k)^2\|H^k\|^2 < 0,$$

and hence

$$\|V^{k+1}\|^2 < \|V^k\|^2. \tag{9}$$

We now prove that $a_0^k \to -\infty$ is impossible. We proceed by contradiction. Assume to the contrary that $a_0^k \to -\infty$ does hold, then $\forall M_2 \geq M_1, \exists K > 0$, such that $a_0^k < -M_2 \leq -M_1$ for $k > K$. Noticing (9), we have $\|V^{k+1}\|^2 < \|V^k\|^2$ when $k > K$, that is, V^k is bounded. So $a_0^k = (V^k)^T H^k$ is also bounded. But this violates the assumption that $a_0^k \to -\infty$. Thus $a_0^k \not\to -\infty$.

The above discussion indicates that $\{a_0^k\}_{k=0}^{\infty}$ has an infinite subsequence that is bounded below. Hence there exist a constant M and a subsequence $\{a_0^{k_n}\}_{n=1}^{\infty}$ such that every a_0^k which satisfies $a_0^k \geq M$ is included in this subsequence. □

Lemma 2 *There exists a constant $M_\varepsilon > 0$ depending on the constant ε in (5), such that $a_0^k \leq M_\varepsilon, \forall k = 1, 2, \cdots$.*

Proof By the weight updating rule, the weight vector W_i^k is refined if and only if $1 - x_0^k = 1 - f(a_0^k) \geq \varepsilon$. Therefore, $a_0^k \leq M_\varepsilon = f^{-1}(1-\varepsilon) > 0$. □

For the weight vector subsequence $\{W_i^{k_n}\}_{n=1}^{\infty}$ corresponding to $\{a_0^{k_n}\}_{n=1}^{\infty}$, we have

Lemma 3 *Assume Case II and (III) (IV), then there exists a constant $C_4 > 0$ such that*

$$A^T W_i^{k_{n+1}} \geq A^T W_i^{k_1} + C_4 n, \forall n = 1, 2, \cdots. \tag{10}$$

Proof Left-multiplying both sides of (6) by A and noticing (4), (5) and (8), we derive

$$A^T W_i^{k+1} = A^T W_i^k + \mu(1-x_0^k)f'(a_i^k)f'(a_0^k)A^T U^k \geq A^T W_i^k + C_3 f'(a_0^k), \tag{11}$$

where $C_3 = \mu\varepsilon C_1 C_2$. If $k \in /\{k_n\}_{n=1}^{\infty}$, because $f'(a_0^k) > 0$, there holds

$$A^T W_i^{k+1} > A^T W_i^k. \tag{12}$$

If $k \in \{k_n\}_{n=1}^{\infty}$, for example $k = k_n$, we conclude from Property 4 and $M \leq a_0^{k_n} \leq M_\varepsilon$ that $f'(a_0^{k_n}) \geq G_{\max\{|M|, M_\varepsilon\}}$. Then (11) implies

$$A^T W_i^{k_n+1} \geq A^T W_i^{k_n} + C_4, \tag{13}$$

where $C_4 = C_3 G_{\max\{|M|, M_\varepsilon\}}$. It follows from (12) and (13) that

$$A^T W_i^{k_{n+1}} > A^T W_i^{k_{n+1}-1} > \cdots > A^T W_i^{k_n+1} \geq A^T W_i^{k_n} + C_4. \tag{14}$$

This immediately results in (10). □

Now, we are in a position to present our main result.

Theorem *Assume (III) and (IV), then the training procedures (6) and (7) converges in finite iteration steps.*

Proof Suppose to the contrary that Case II is right. Then $\{W_{i_0}^{k_n}\}_{n=1}^{\infty}$ satisfies (10) and $\|W_i^{k_n}\| \leq C$ (C is the constant in (IV)). By the Schwartz inequality, there holds

$$\|A\| \geq \frac{A^T W_{i_0}^{k_{n+1}}}{\|W_{i_0}^{k_{n+1}}\|} \geq \frac{A^T W_{i_0}^{k_1} + C_4 n}{C} \to \infty, n \to \infty,$$

leading to a contradiction. So Case I must be true, that is, the online BP algorithms (6) and (7) must converge in finite number of iteration steps. □

Remark 1 The assumption $v_i \in R^+$: As pointed out in [1], $v_i \in R^+$ can be reasonably obtained by an inversion of the sign of W_i.

Remark 2 The boundedness of $\|W_i\|$: This assumption is restrictive in theory, but is naturally adopted in practice and numerical experiment. Usually the larger values of $\|W_i\|$ are difficult or expensive to implement by hardware. So when $\|W_i\| \to \infty$, we normally terminate the process of the algorithm and start from another initial value W_i^0.

Appendix: An error in [1]

The following estimate ((17) in Lemma 2 of [1]) plays a central role in the proof in [1]:

$$\|W_i(K)\|^2 \leq \|W_i(0)\|^2 + \sum_{k=0}^{K-1} \mu_i^2(W(k), k_{\mathrm{mod}Q}) \cdot y_i^2(W(k), V(k), k_{\mathrm{mod}Q}) \|U_b(k_{\mathrm{mod}Q})\|^2.$$

But this is not correct. Actually, based on the triangular inequality, for any a, b and $c \in R^m$, if $a = b + c$, then $\|a\| \leq \|b\| + \|c\|$. So using (9) in [1]

$$W_i(k+1) = W_i(k) - \mu_i(W(k), k_{\mathrm{mod}Q}) \cdot y_i(W(k), V(k), k_{\mathrm{mod}Q}) U_b(W(k), k_{\mathrm{mod}Q}),$$

we can only obtain for instance

$$\|W_i(K)\| \leq \|W_i(0)\| + \sum_{k=0}^{K-1} \mu_i(W(k), k_{\mathrm{mod}Q}) \cdot |y_i(W(k), V(k), k_{\mathrm{mod}Q})| \|U_b(k_{\mathrm{mod}Q})\|,$$

but not (17) of [1]. There are no obvious ways to correct this error in the framework of [1].

References:

[1] GORI M, MAGGINI M. *Optimal convergence of on-line backpropagation* [J]. IEEE Tran. Neural Networks, Volume: 7 , Issue: 1, 1996, 251–154.

[2] ROSENBLATT F. *Principles of Neurodynamics* [M]. Spartan, New York, 1962.
[3] WIDROW B, HOFF M E. *Adaptive Switching Circuits* [M]. in J.A. Anderson & E. Rosenfeld, Neurocomputing: foundations of research, The MIT Press, Cambridge, MA, 1988.
[4] WU Wei, SHAO Z. *Convrgence of Online Gradient Methods for Continuous perceptrons with Linearly Separable Trainin Patterns* [J]. Appl. Math. Lett., Volume: 16, Issue: 7, October, 2003, pp. 999-1002.

具线性可分训练样本时在线 BP 神经网络的有限收敛性

邵郅邛, 吴微, 杨洁
(大连理工大学应用数学系, 辽宁 大连 116023)

摘要: 当训练样本线性可分时, 本文证明前馈神经网络的在线 BP 算法是有限次收敛的.

关键词: 非线性前馈神经网络; 在线 BP 算法; 有限收敛性; 线性可分训练样本.