Convergence Properties of an Online Learning Algorithm in Neural Network Models of Complex Systems

V.N. Azarskov, S.A. Nikolaienko
Aircraft Control Systems Dept.
National Aviation University
Kyiv, Ukraine
azarskov@nau.edu.ua
s_nicolaenko@ukr.net

L.S. Zhiteckii
Int. Centre of Inform. Tech. & Syst.
Institute of Cybernetics
Kyiv, Ukraine
leonid_zhiteckii@i.ua

Abstract—**Asymptotic behavior of the online gradient algorithm with a constant step size employed for learning in neural network models of nonlinear systems having hidden layer are studied. The sufficient conditions guaranteeing the convergence of this algorithm in the random environment are established.**

Keywords—nonlinear model; neural network; gradient algorithm; learning; convergence

I. Introduction

Neural networks containing at least one hidden layer play a role of universal models for any reasonable complex nonlinear systems, in particular, flight control systems. This fact motivates the theoretical studies of learning algorithms for the neural network models. Significant breakthrough in this research area has been achieved in recent works [1]–[8]. Namely, the convergence results have been derived in [7] provided that input signals have a probabilistic nature. In their stochastic approach, the learning rate goes to zero as the learning process tends to infinity. Unfortunately, this gives that the learning goes faster in the beginning and slows down in the late stage.

The convergence analysis of learning algorithms with deterministic (non-stochastic) nature has been given in [8] by assuming that the learning set is finite. The difficulties in establishing the convergence results are that the neural networks contain the parameters which appear nonlinearly in their equations. To the best of author's knowledge, there are no results in literature concerning the convergence properties of training procedures with a fixed step size applicable to the case of infinite learning set.

This paper generalizes some results obtained by the authors in [9]. The main effort is focused on establishing sufficient conditions under which the online gradient algorithms applied for sequential learning neural networks with a constant step size will converge in the case of infinite learning set. The key idea in studying their asymptotic properties is based on exploiting the stochastic counterpart of the Lyapunov function method, which is known in the probability theory as the supermartingale.

II. Problem Formulation

Let

$$y = F(x) \tag{1}$$

be some nonlinear unknown function describing a complex system. In this equation, $y \in \mathrm{IR}$ and $x \in \mathrm{IR}^N$ are the output scalar and input vector variables, respectively, available for the measurement at each nth time instant $(n = 1, 2, \ldots)$. This implies that

$$y(n) = F(x(n-1)) \tag{2}$$

with an unknown mapping $F: \mathrm{IR}^N \to \mathrm{IR}$.

To approximate (1), the two-layer neural network containing M $(M \geq 1)$ neurons in its hidden layer is employed. The inputs to the each jth neuron of this layer at the time instant n are the components of $x(n-1)$. Its output signal at the nth time instant is given by

$$y_j^{(1)}(n) = \sigma\left(b_j^{(1)} + \sum_{i=1}^{N} w_{ij}^{(1)} x_i(n-1)\right), \quad j = 1, \ldots, M, \tag{3}$$

where $x_i(n-1)$ denotes the ith component of $x(n-1)$, and $w_{ij}^{(1)}$ and $b_j^{(1)}$ are the weight coefficients and the bias of this jth neuron, respectively. $\sigma(\cdot)$ represents the so-called activation function. There is only one neuron in the output (second) layer, whose inputs are the outputs of the hidden layer's neurons. The output signal of second layer, $y^{(2)}(n)$, at the time instant n is determined by

978-1-4799-3305-1 © 2013 IEEE

$$y^{(2)}(n) = \sum_{j=1}^{M} w_j^{(2)} y_j^{(1)}(n) + b^{(2)}, \quad (4)$$

where $w_1^{(2)},\ldots,w_M^{(2)}$ are the weights of this neuron and $b^{(2)}$ is its bias.

Since $\sigma(\cdot)$ is assumed to be nonlinear, it follows from (3), (4) together with (2) that $y^{(2)}(n)$ is a nonlinear function depending on $x(n-1)$ and also on the $(M(N+2)+1)$-dimensional parameter vector

$$w = [w_{11}^{(1)},\ldots,w_{N1}^{(1)}, b_1^{(1)},\ldots,w_{1M}^{(1)},\ldots,w_{NM}^{(1)}, b_M^{(1)} : w_1^{(2)},\ldots,w_M^{(2)}, b^{(2)}]^T.$$

To emphasize this fact, define the output signal of the neural network in the form

$$y^{(2)}(n) = \mathrm{NN}(x(n-1), w) \quad (5)$$

with $\mathrm{NN}: \mathrm{IR}^N \times \mathrm{IR}^{M(N+2)+1} \to \mathrm{IR}$.

The following basic assumption is made. There exists at least an unique $w = w^* \in \mathrm{IR}^{M(N+2)+1}$ such that $F(x)$ can explicitly be approximated by $\mathrm{NN}(x, w^*)$ in the sense of

$$F(x) \equiv \mathrm{NN}(x, w^*) \quad (6)$$

for all x from a given compact set $X \subset \mathrm{IR}^N$.

Define the training sequence $\{(x(n-1), y(n))\}_{n=1}^{\infty}$ of the measurable pairs in which $x(n-1)$s are taken randomly from X. Then, the online learning algorithm for updating the parameter estimate $w(n)$ is formulated as the following standard recursive gradient procedure:

$$w(n) = w(n-1) + \eta\, \tilde{e}(n, w(n-1))\, \mathrm{grad}_w \mathrm{NN}(x(n-1), w(n-1)). \quad (7)$$

In this algorithm,

$$\tilde{e}(n, w(n-1)) = y(n) - \mathrm{NN}(x(n-1), w(n-1)) \quad (8)$$

is the current estimation error and $\mathrm{grad}_w \mathrm{NN}(x(n-1), w(n-1))$ *denotes the gradient of* $\mathrm{NN}(x, w)$ *at the point* $w = w(n-1)$, *and* $\eta \equiv const > 0$ *is its step size (the learning rate).*

The problem is to study the properties of sequence $\{w(n)\}$ caused by (7), (8) as n tends to ∞.

Equations (2) and (7) together with (5), (6) and (8) describe the closed-loop system for adaptive identification of (1).

III. Convergence Analysis

A. Preliminaries

To analyze the asymptotic behavior of (7), (8), the scalar non-negative function $V(w)$ given by

$$V(w) = 0 \text{ for } w = w^*, \qquad V(w) > 0 \text{ for } w \neq w^* \quad (9)$$

is exploited.

The variable $V_n := V(w(n))$ becomes immediately the Lyapunov function of the algorithm (7), (8) if only

$$V_n \leq V_{n-1} \qquad \forall n. \quad (10)$$

Since $V_n \geq 0$, the condition (10) under which V_n does not increase is sufficient for existing a limit

$$\lim_{n\to\infty} V_n = V_\infty, \quad (11)$$

where V_∞ is a random number depending on $w(0)$ and $\{x(n)\}$.

In the presence of the one-point $W^* = \{w^*\}$, the function $V(w)$ satisfying (9) is usually chosen as

$$V(w) = \left\| w^* - w \right\|^2, \quad (12)$$

where $\|\cdot\|$ denotes the Euclidean vector norm. It turned out that if the neural network contains the hidden layer, then W^* consists of several isolated w^*s. In particular, in the simplest case, when there is one neuron in the hidden layer $(N = 1,\ M = 1)$ and $\sigma(\cdot)$ is described by

$$\sigma(s) = \frac{1}{1+\exp(-s)}, \quad (13)$$

the set W^* contains two points: $\mathbf{w}_1^* = [w_1^*, w_2^*, w_3^*, w_4^*]^T$ and $\mathbf{w}_2^* = [-w_1^*, -w_2^*, -w_3^*, w_3^* + w_4^*]^T$.

In the case when W^* is not one-point, $V(w)$ is designed as

$$V(w) = \inf_{\mathbf{w}^* \in W^*} \left\| w^* - w \right\|^2 \quad (14)$$

but not as defined in (12).

We first observed in simulation examples that $\{w(n)\}$ may not converge even in the presence of bounded $\{x(n)\}$ if there are no additional restrictions on this input sequence. Such an ultimate feature of (7), (8) implies that

$$\lim_{n\to\infty} w(n) = w_\infty \tag{15}$$

may not exist. Nevertheless, if (15) is achieved, then the following asymptotic properties of $\{w(n)\}$ can be established:

a) $\{w(n)\}$ converges to some w_∞ in sense of (15) with

$$w_\infty \in \lim\inf W_n,$$

where

$$\lim\inf W_n := \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} W_k$$

denotes the limit set in which

$$W_n := \{w:\ y(n) - \mathrm{NN}(x(n-1), w) = 0\};$$

b) the identification error given by (8) satisfies

$$\lim_{n\to\infty} \tilde{e}(n, w(n-1)) = 0. \tag{16}$$

Note that $\lim\inf W_n$ represents a nonlinear manifold on $\mathrm{IR}^{M(N+2)+1}$ whose dimension satisfies

$$0 \le \dim \lim\inf W_n \le M(N+2).$$

B. Simulation Examples

To demonstrate these asymptotic properties, three simulation experiments with

$$y = \frac{3.75 + 0.05\exp(-7.15x)}{1 + 0.19\exp(-7.15x)}$$

was performed. This nonlinear function can explicitly be approximated by the two-layer neural network described by (5), (6), (8) and (13) with $w_{11}^{(1)*} = 7.15$, $b_1^{(1)*} = 1.65$, $w_1^{(2)*} = 3.45$, $b^{(2)*} = 0.3$. In all of these experiments, η was taken as $\eta = 0.02$.

The simulation results are depicted in Figures 1-3.

Fig. 1 shows that $\{V_n\}$ has no limit if the input sequence $\{x(n)\}$ is non-stochastic. (The definition of the non-stochastic sequence can be found in the paper [9].) In this case, the model error $\tilde{e}(n, w(n-1))$ does not go to zero, i. e., (16) is not satisfied.

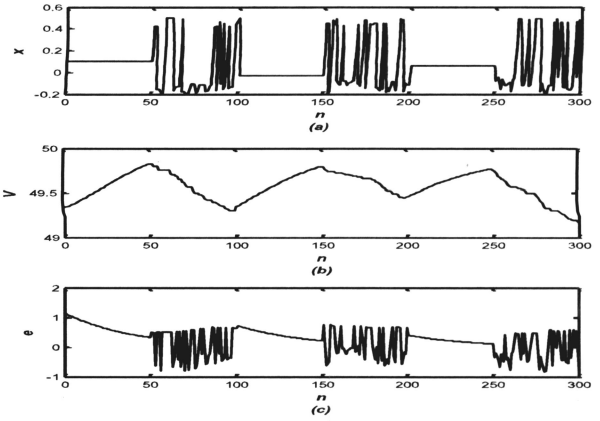

Fig. 1. Learning processes in simulation experiment 1: (a) the input signal; (b) the function V_n given by (14); (c) the current model error

In second experiment, $x(n)$s were sequentially chosen from the finite set containing three points: $x^{(1)} = -0.4442$; $x^{(2)} = 0.5158$; $x^{(3)} = 0.8761$. Fig. 2 illustrates the result of this experiment with initial $w_{11}^{(1)}(0) = 0.529$, $b_1^{(1)}(0) = -0.5012$, $w_1^{(2)}(0) = -0.9168$, $b^{(2)}(0) = 1.0409$.

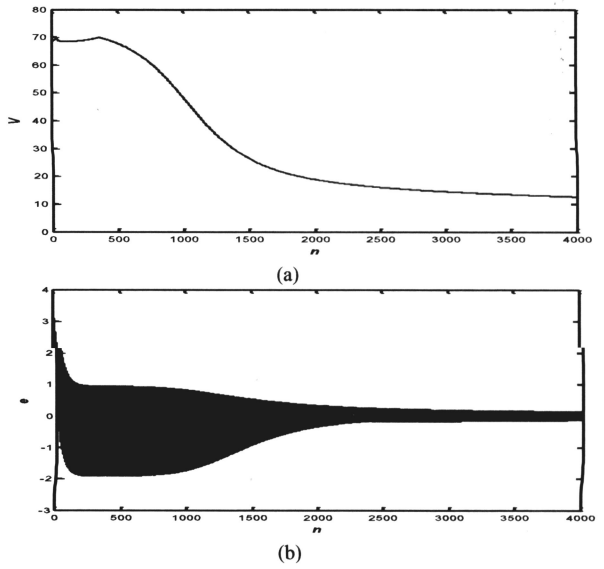

Fig. 2. Learning processes in simulation experiment 2: (a) the function V_n given by (14); (b) the current model error

We can observe that $\{V_n\}$ is convergent, however its convergence is not monotonic as in (10). It turned out that in this case, $\{w(n)\}$ converges to $w_\infty = [5.4120,\ 1.3172,\ 3.8233,\ -0.0475]^T$ which lies on $\lim\inf W_n$ but not to one of two points $w_1^* = [7.15,\ 1.65,\ 3.45,\ 0.3]^T$ or to $w_1^* = [-7.15,\ -1.65,\ -3.45,\ 3.75]^T$.

The case where $\{V_n\}$ converges monotonically is demonstrated in Fig. 3. In this case, the initial estimates were chosen as follows: $w_{11}^{(1)}(0) = 1.4$, $b_1^{(1)}(0) = -0.1$, $w_1^{(2)}(0) = -0.56$, $b^{(2)}(0) = 0.46$.

It turned out that $\{w(n)\}$ tends to the limit point $w_1^* = [7.15, 1.65, 3.45, 0.3]^T$ as $n \to \infty$.

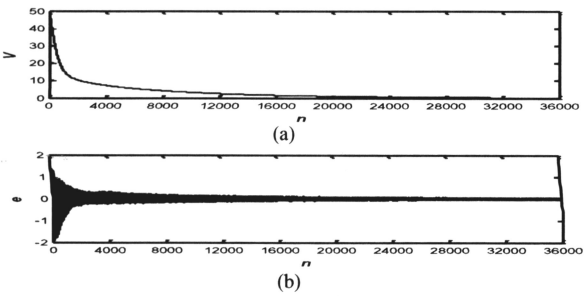

Fig. 3. Learning processes in simulation experiment 3:
(a) the function V_n given by (14); (b) the current model error

IV. Main Result

Main theoretical result concerning the asymptotical behavior of (7), (8) is based on following additional assumptions:

$\{x_i(n)\}$ are the stochastic sequences of independent random variables having the probability density function

$$p(x(n) \mid x(n-1), \ldots, x(0)) \equiv p(x(n)) := p(x) \qquad (17)$$

with the properties that

$$P\{x(n) \in X'\} = \int_{x \in X'} p(x)\, dx > 0, \qquad (18)$$

for any subset $X' \subset X$, and

$$P\{x(n) \in X''\} = 0 \qquad (19)$$

if $\dim X'' = 0$, where $P\{\cdot\}$ denotes the probability of corresponding event.

Let $W(w^*)$ denote a neighborhood of some $w^* \in W^*$ which does not contain another points of W^*. With this $W(w^*)$, we have established that if the assumptions (6), (17) – (19) are satisfied and the conditions

$$0 < \eta < 2,$$

$$\int_{x \in X} [\mathrm{NN}(x, w^*) - \mathrm{NN}(x, w)] \mathrm{grad}_w^T \mathrm{NN}(x, w)(w^* - w) p(x)\, dx$$
$$\geq \int_{x \in X} [\mathrm{NN}(x, w^*) - \mathrm{NN}(x, w)]^2 \left\| \mathrm{grad}_w \mathrm{NN}(x, w) \right\|^2 p(x)\, dx$$

hold for any $x \in X$ *and for arbitrary* w *from* $W(w^*)$, then the limit (11) is valid with probability 1. Again,

$$\lim_{n \to \infty} w(n) = w^*$$

almost sure (a. s.).

The proof of this result essentially utilizes the Borel-Cantelli lemma and *Doob's martingale convergence* theorem (see [9]).

V. Conclusion

In general case, the standard online gradient algorithms applied to sequential learning in neural networks with hidden layer may not converge. To guarantee their convergence, certain conditions need to be satisfied.

References

[1] L. Behera, S. Kumar, and A. Patnaik, "On adaptive learning rate that guarantees convergence in feedforward networks," IEEE Trans. on Neural Networks, vol. 17, pp. 1116–1125, May 2006.

[2] 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, pp. 469–476, 2007.

[3] V. Tadic and S. Stankovic, "Learning in neural networks by normalized stochastic gradient algorithm: Local convergence," in Proc. 5th Seminar Neural Netw. Appl. Electr. Eng, Yugoslavia, Sept. 2000, pp. 11–17.

[4] Wu, W., G. Feng, and X. Li, "Training multilayer perceptrons via minimization of ridge functions," Advances in Comput. Mathemati vol. 17, pp. 331–347, 2002.

[5] W. Wu, G. Feng, X. Li, and Y. Xu, "Deterministic convergence of online gradient method for BP neural networks," IEEE Trans. on Ne Networks, vol. 16, pp. 1–9, March 2005.

[6] N. Zhang, W. Wu, and G. Zheng, "Convergence of gradient method momentum for two-layer feedforward neural networks," IEEE Trans Neural Networks, vol. 17, pp. 522–525, February 2006.

[7] H. Zhang, W. Wu, F. Liu, and M. Yao, "Boundedness and conve of online gradient method with penalty for feedforward neur works," IEEE Trans. Neural Networks, vol. 20, pp. 1050–105 2009.

[8] Z. Xu, R. Zhang, and W. Jing "When does online BP trainin verge?" IEEE Trans. Neural Networks, vol. 20, pp. 1529–1539, ber 2009.

[9] L.S. Zhiteckii, V.N. Azarskov, and S.A. Nikolaienko, "Conver learning algorithms in neural networks for adaptive identific nonlinearly parameterized systems," in Proc. 16th IFAC Symp System Identification, Brussels, Belgium, July 10 – 13, 2012, p 1598.