UDC 681.5

CONVERGENCE OF SEQUENTIAL GRADIENT LEARNING ALGORITHMS IN NEURAL NETWORKS FOR ONLINE IDENTIFICATION OF NONLINEAR SYSTEMS: A SPECIAL CASE

L.S. Zhiteckii, S.A. Nikolaienko

International Research and Training Center for Information Technologies and Systems

leonid_zhiteckii@i.ua

Стаття стосується асимптотичних властивостей деякої процедури навчання в реальному часі для ідентифікації нелінійних систем з використанням нейронних мереж як моделей цих систем. Представлені умови ймовірносної збіжності цієї процедури для спеціального випадку, коли нелінійність може бути точно апроксимована належною нейронною мережею.
Ключові слова: ідентифікація, нелінійна система, нейронна мережа, алгоритм навчання, стохастичне середовище, збіжність.

The paper deals with the asymptotic properties of an online learning procedure for identifying nonlinear systems via neural networks models of these systems. The probabilistic convergence conditions of this procedure are presented for the special case where a nonlinearity can exactly be approximated by a suitable neural network.
Keywords: identification, nonlinear system, neural network, learning algorithm, stochastic environment, convergence.

Статья касается асимптотических свойств некоторой процедуры обучения в реальном времени для идентификации нелинейных систем с использованием нейронных сетей в качестве моделей этих систем. Представлении условия вероятностной сходимости этой процедуры для специального случая, когда нелинейность может быть точно аппроксимирована подходящей нейронной сетью.
Ключевые слова: идентификация, нелинейная система, нейронная сеть, алгоритм обучения, стохастическая среда, сходимость.

Introduction. The problem of identifying complex unknown systems in the presence of noise remains important from both theoretical and practical point of view up to now. Significant progress in this research area were achieved in the frameworks of well-known group method of data handling (GMDH) advanced by A. G. Ivakhnenko in the late 1960s to deal with a finite set of training examples to be used for deriving mathematical models of unknown systems [1]. Over the past decades, interest has been increasing toward the use of multilayer neural networks as models for the adaptive identification of nonlinearly parameterized dynamic systems [2–5]. Several learning methods for updating the weights of neural networks have been advanced in literature. Most of these methods rely on the gradient concept [5, 6]. Although this concept has been successfully used in many empirical studies, there are very few fundamental results dealing with the convergence of gradient algorithms for learning neural networks. One of these results is based on utilizing the Lyapunov stability theory [3, 6].

The asymptotic behavior of online adaptive gradient algorithms for the network learning has been studied by many authors [7–22]. In particular, the convergence of the learning process for the so-called feedforward network models with single hidden layer is investigated in [7] by using the stochastic approximation theory. The convergence results have been derived in [9–15] among many others 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 algorithm with deterministic (non-stochastic) nature has been given in [16–21]. In contrast to the stochastic approach, several of these results allow to employ a constant learning rate [18, 22]. However, they assume that learning set must be finite whereas in online identification schemes, this set is theoretically infinite. To the best of author's knowledge, there are no general results in literature concerning the global convergence properties of training procedures with a fixed learning rate applicable to the case of infinite learning set.

The distinguishing feature of multi-layer neural networks is that they describe some nonlinearly parameterized models needed to be identified. This leads to difficulties in deriving their convergence properties for a general case. To avoid these difficulties in non-stochastic case, the assumption that similar nonlinear functions need to be convex (concave) is introduced in [23]. However, such an assumption is not appropriate for neural network's description of nonlinearity.

A popular approach to analyze the asymptotic behavior of online gradient algorithms in stochastic case is based on Martingale convergence theory [24]. This approach has been exploited in [25, 26] to derive some local convergence in stochastic framework for standard online gradient algorithms with the constant learning rate.

This paper is an extension of [25, 26]. The main efforts is focused on establishing sufficient conditions under which the global convergence of gradient algorithm for learning neural networks models in the stochastic environments will be achieved. The key idea in deriving these convergence results is based on the use of the Lyapunov methodology [27].

1. System identification using a neural network model

Let

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

be the nonlinear equation in the compact form describing a complex system to be identified. In this equation, $y(n) \in \mathrm{IR}$ and $x(n) \in \mathrm{IR}^N$ are the scalar output and the so-called state vector, respectively, available for the measurement at each nth time instant, $\xi(n)$ is noise at some time instant $(n = 1, 2, \ldots)$, and $F: \mathrm{IR}^N \to \mathrm{IR}$ represents some unknown nonlinear mapping. (Note that $x(n)$ may include the current inputs of this system and possibly its past inputs and also outputs; see [6, sect.

5.15].) Without loss of generality, one supposes that the nonlinearity $F(x)$ is the continuous and smooth function on a bounded set $X \subset \mathrm{IR}^N$ (diam $X < \infty$).

To approximate $F(x)$ by a suitable nonlinearly parameterized function, the two-layer neural network model containing M $(M \geq 1)$ neurons in its hidden layer is employed. The inputs to the each *j*th neuron of this layer at the time instant n are the components of $x(n)$. Its output signal at the *n*th time instant is specified as

$$y_j^{(1)}(n) = \sigma\left(b_j^{(1)} + \sum_{i=1}^{N} w_{ij}^{(1)} x_i(n) \right), \quad j = 1, \ldots, M, \tag{2}$$

where $x_i(n)$ denotes the *i*th component of $x(n)$, and $w_{ij}^{(1)}$ and $b_j^{(1)}$ are the weight coefficients and the bias of this *j*th neuron, respectively. $\sigma(\cdot)$ denotes the so-called activation function defined usually as the sigmoid functions

$$\sigma(s) = \frac{1}{1 + \exp(-s)} \tag{3}$$

or

$$\sigma(s) = \tanh\ (s). \tag{4}$$

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

$$y^{(2)}(n) = \sum_{j=1}^{M} w_j^{(2)} y_j^{(1)}(n) + b^{(2)}, \tag{5}$$

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

Since $\sigma(\cdot)$s defined by (3) and (4) are nonlinear, it follows from (2), (5) that $y^{(2)}(n)$ is the nonlinear function depending on $x(n-1)$ and also on the $(M(N+2)+1)$-dimensional parameter vector

$$w = [w_{11}^{(1)}, \ldots, w_{N1}^{(1)}, b_1^{(1)}, \ldots, w_{1M}^{(1)}, \ldots, w_{NM}^{(1)}, b_M^{(1)} \vdots w_1^{(2)}, \ldots, w_M^{(2)}, b^{(2)}]^T. \tag{6}$$

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

$$y^{(2)}(n) = \mathrm{NN}(x(n), w) \tag{7}$$

using the notation $\mathrm{NN}: \mathrm{IR}^N \times \mathrm{IR}^{M(N+2)+1} \to \mathrm{IR}$. Taking into account that the neural network plays the role of a model of the nonlinearity $F(x)$, rewrite (7) as follows:

$$y_{\mathrm{mod}}(n) = \mathrm{NN}(x(n), w). \tag{8}$$

Optimal value $w = w^*$ specified by the least modulus

$$w^* = \arg \min_{w} \max_{x \in X} |F(x) - \mathrm{NN}(x, w)| \tag{9}$$

and also the discrepancy

$$e = F(x) - \mathrm{NN}(x, w)$$

between $F(x)$ and the output of its neural network's model for a fixed w corresponding to (8) are unknown.

To do an adaptation of the neural network model to the uncertain system (1), the standard online gradient learning algorithm

$$w(n) = w(n-1) - \eta(n)\nabla_w Q(x(n), w(n-1)) \tag{10}$$

taken, for example, from [5,6] is usually utilized. In this algorithm, $\nabla_w Q(x(n), w(n-1))$ represents the gradient of the quadratic loss function

$$Q(x, w) = \frac{1}{2}[y - \mathrm{NN}(x, w)]^2 \tag{11}$$

with respect to w at $w = w(n-1)$ for given $x = x(n)$, and $\eta(n)$ is the learning rate (step size) of (10). Due to (11) we have

$$Q(x(n), w(n-1) = \frac{1}{2}[y(n) - \mathrm{NN}(x(n), w(n-1))]^2 \tag{12}$$

with the variable

$$e(n, w(n-1)) = y(n) - \mathrm{NN}(x(n), w(n-1)) \tag{13}$$

representing the current model error which can be measured at the *n*th time instant. Now, using (11) – (13), rewrite the learning algorithm (10) as follows:

$$w(n) = w(n-1) + \eta(n)\, e(n, w(n-1))\, \nabla_w \mathrm{NN}(x(n-1), w(n-1)). \tag{14}$$

Thus, (2), (5), (7) and (14) describe the learning system necessary for the adaptive identification of (1). For better understanding its performance, the structure of this system is depicted in Fig. 1.

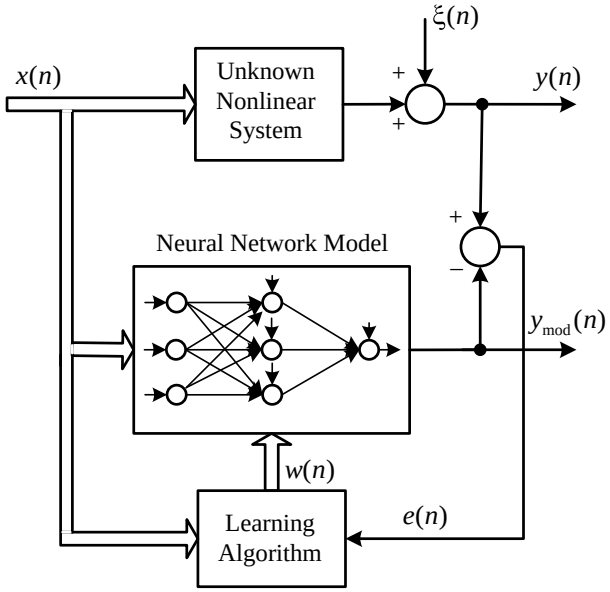


Fig. 1. Configuration of online learning system

2. Statement of the problem

Consider a special case where $F(x)$ can exactly be approximated by a neural network representation for all $x \in X$ implying

$$F(x) \equiv \mathrm{NN}(x, w^*). \tag{15}$$

In this case called in [5, p. 304] as the ideal case, one has $e(n, w^*) \equiv 0$ with w^* given by (9) if only $\xi(n)$ is absent. Note that this special case is similar to the so-called the hypothesis of representation [6, p. 81] advanced by M.A. Aizerman, E.M. Braverman and L.I. Rozonoer in the machine learning theory at the beginning 1960s.

Suppose $\{x(n)\}$ is an infinite sequence of vectors belonging to the bounded X.

The aim of this paper consists in studying the asymptotic properties of the learning procedure (14) caused by this $\{x(n)\}$. More certainty, the following problem is stated. It is required to derive the conditions under which $\{w(n)\}$ will converge in the sense that

$$\lim_{n\to\infty} w(n) = w_\infty \qquad \text{with } \| w_\infty \| < \infty. \tag{16}$$

3. Preliminaries

First, recall that the condition

$$\eta(n)\, e(n, w(n-1))\, \nabla_w \mathrm{NN}(x(n-1), w(n-1)) \xrightarrow[n\to\infty]{} 0 \tag{17}$$

followed from (14) is necessary to achieve the limit (16), for a given $\{x(n)\}$ [6, sect. 3.13]. Since $\nabla_w \mathrm{NN}(x(n), w(n)) \not\equiv 0$, it can be observed that or the condition 1 of the form

$$\eta(n) \equiv \mathrm{const}, \qquad e(n, w(n)) \to 0 \quad \text{as} \quad n \to \infty$$

or the condition 2 of the form

$$\eta(n) \to 0, \qquad e(n, w(n)) \not\to 0 \quad \text{as} \quad n \to \infty$$

are required to satisfy (17). Note that the condition 1 cannot take place if the noise $\xi(n)$ are present because $e(n, w^*) \equiv \xi(n)$ (due to (1), (13), (15)).

It turned out that in the special case, the set W^*, containing these w^*s becomes not one-point [25, 26]. To show it, put $N = 1$, $M = 1$. Due to (6), this implies $w^* \in \mathrm{IR}^4$. Let $w^* = [w_1^*, w_2^*, w_3^*, w_4^*]^T$ be a vector satisfying (15). Then, (2) and (5) together with (3) give that another $w^* = [-w_1^*, -w_2^*, -w_3^*, w_3^* + w_4^*]^T$ will also satisfy the equality (15).

Introduce the scalar variable $\| w^* - w \|^2$ representing the square of Euclidean distance between w and a w^*, and define

$$V(w) = \inf_{w^* \in W^*} \| w^* - w \|^2 . \tag{18}$$

Denote $V_n := V(w(n))$. Since $V_n \geq 0$ (due to (18)), it is clear that if

$$V_n \leq V_{n-1} \tag{19}$$

then the sequence $\{V_n\} := V_0,...,V_n,...$ has always a limit, V_∞, as n tends to infinity, i.e.,

$$\lim_{n\to\infty} V_n = V_\infty, \tag{20}$$

meaning that the algorithm (14) converges. On the other hand, the fact that $\{V_n\}$ is monotonical non-increasing sequence is not necessary to achieve (20) in principle.
Note that the existence of the limit (20) does not imply that $V_\infty = 0$ even when the condition (15) is satisfied. Moreover, this limit may not exist if $\{x(n)\}$ is an arbitrary sequence leading to the violation of (19) [25]. Nevertheless, if the asymptotic property (16) takes place, then $\{w(n)\}$ converges to some $w_\infty \in \lim\ \inf\ W_n$ where

$$\lim\ \inf\ W_n := \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} W_k \tag{21}$$

denotes the so-called limit set introduced in [24, sect. 1.3] in which $W_n := \{w:\ y(n) - \mathrm{NN}(x(n-1), w) = 0\}$.

Note that the limit set, $\lim\ \inf\ W_n$, given by (21) represents a nonlinear manifold on $\mathrm{IR}^{M(N+2)+1}$ whose dimension satisfies $0 \leq \dim\ \lim\ \inf\ W_n \leq M(N+2)$.

It can be understood that the algorithm (14) "attempts" to solve the infinite set of the equations

$$y(n) - \mathrm{NN}(x(n-1), w) = 0,\ \ n = 1, 2,\ldots \tag{22}$$

with respect to unknown $w \in \mathrm{IR}^{M(N+2)+1}$. In fact, this algorithm may give the solution $w = w_\infty$ of the remainder of (22), which is determined as the limit set (21) but not as W^*.

It was observed that the condition (19) meaning that $\{V_n\}$ is the monotonically non-increasing sequence may not be satisfied if the neural network model contains the hidden layer, in general.

To demonstrate some asymptotic properties of (14), two simulation experiments with the scalar nonlinear system (1) having the nonlinearity

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

were conducted. It can be shown that this nonlinearity can explicitly be approximated by the two-layer neural network model described by (2), (3), (5) and (7) with the components of two $w = w^{*(1)}$, $w = w^{*(2)}$ summarized in Table 1.

Table 1

Parameters of neural network model

Exp. No	Parameter	$w_{11}^{(1)}$	$b_1^{(1)}$	$w_1^{(2)}$	$b^{(2)}$
1, 2	Components of $w^{*(1)}$	7,15	1.65	3.45	0.3
	Components of $w^{*(2)}$	-7.15	-1.65	-3.45	3.75
1	Initial estimate	0.53	-0.50	-0.92	1.04
	Final estimate	5.41	1.32	3.82	-0.05
2	Initial estimate	0.38	-0.57	-0.98	1.14
	Final estimate	-5.13	-1.52	-4.20	3.78

In all of the experiments, $\eta(n)$ was taken as $\eta(n) \equiv \eta = 0.01$. In these experiments, $\{x(n)\}$ was generated as sequence of independent identically distributed (i.i.d.) pseudo random numbers on $X = [-1.0, 1.0]$. The duration of the learning processes was always equal to 40 000 steps.

Simulation results of first and second experiments are presented in Fig. 2 left and right, respectively. The initial estimated $w(0)$ in both examples was chosen so that the distance between $w(0)$ and W^* was large enough, and the condition $V^{(1)}(w(0)) < V^{(2)}(w(0))$ was satisfied. It was observed that at an initial stage of the learning process, $\{V_n^{(1)}\}$ was increasing and $V_n^{(1)} > V_n^{(2)}$ for several $n = 1, 2, \ldots,$ as shown in Fig. 2, left. Further, $\{V_n^{(1)}\}$ became decreasing. Such a behavior of these sequence leaded to appearing the feature that $V_n^{(1)} < V_n^{(2)}$ for all sufficiently large n.

In the second example, the initial $w(0)$ was chosen to be close to that in the first example. One can observe that in this case, $V_n \equiv V_n^{(1)}$ (see Fig. 2, right).

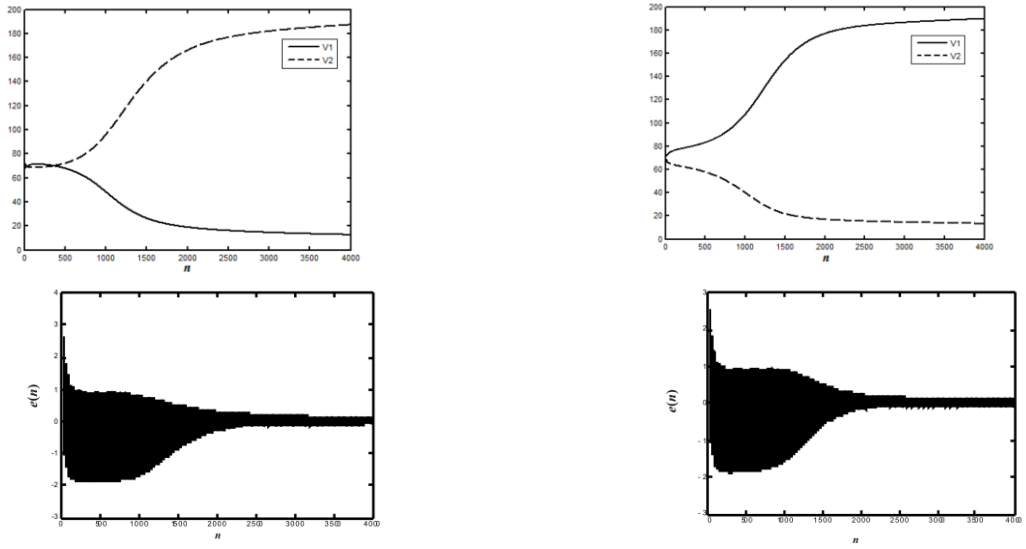


Fig. 2. Behavior of gradient learning algorithm (14) in Examples 1 (left) and 2 (right) in the absence of noise

It turned out that in there simulation examples, the condition (19) is not satisfied whereas the learning algorithm (14) remains indeed convergent.

Thus, additional assumptions with respect to $\{x(n)\}$ are required to guarantee the convergence of $\{w(n)\}$.

4. Local and global convergence results

Assumption 1. $\{x(n)\}$ is a sequence of vectors appearing randomly in accordance with some probability density function $p(x)$ such that

$$\int_X p(x)\,dx = 1.$$

Furthermore, $p(x)$ has the following properties:

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

for any subset $X' \subset X$ whose dimension is N, and

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

if $\dim X'' < N$, where $P\{\cdot\}$ denotes the probability of corresponding random event.

Assumption 2. It is assumed that $p(x)$ represents a continuous function which may become zero only at some isolated points on X.

Assumption 3. The noise is absent, i.e., $\xi(n) \equiv 0$. In this case, $Q(x, w)$ defined in (11) becomes

$$Q(x, w) = \frac{1}{2}[F(x) - \mathrm{NN}(x, w)]^2. \tag{23}$$

Introduce the performance index

$$J(w) = E\{Q(x, w)\} \tag{24}$$

which evaluates the quality of learning process with $Q(x, w)$ given by (23). In this expression,

$$E\{Q(x, w)\} := \int_X [F(x) - \mathrm{NN}(x, w)]^2\, p(x)\,dx$$

denotes the expectation of $Q(x, w)$ with respect to the random xs.

Let $W_\varepsilon(w^*)$ denote an ε-neighborhood of some $w^* \in W^*$ defined as $W_\varepsilon(w^*) := \{w : \| w^* - w \| < \varepsilon\}$, which does not contain another points of W^*. Suppose

a) the assumption 1 – 3 are valid;

b) the condition

$$\int_{x \in W_\varepsilon(w^*)} [\mathrm{NN}(x, w^*) - \mathrm{NN}(x, w)] \nabla_w^T \mathrm{NN}(x, w)(w^* - w) p(x)\,dx$$

$$\geq \int_{x \in W_\varepsilon(w^*)} [\mathrm{NN}(x, w^*) - \mathrm{NN}(x, w)]^2 \left\| \nabla_w \mathrm{NN}(x, w) \right\|^2 p(x)\, dx \qquad (25)$$

meaning

$$E\{[\mathrm{NN}(x, w^*) - \mathrm{NN}(x, w)] \nabla_w^T \mathrm{NN}(x, w)(w^* - w)\}$$
$$\geq E\{[\mathrm{NN}(x, w^*) - \mathrm{NN}(x, w)]^2 \left\| \nabla_w \mathrm{NN}(x, w) \right\|^2\}$$

are satisfied for all $x \in X$ and for any w, w^* from $\mathrm{IR}^{M(N+2)+1}$;

c) an initial $w(0)$ satisfies $w(0) \in W_\varepsilon(w^*)$.

In the work [25] it has been established that, under the conditions a) – c), the limit of $\{w(n)\}$ exists almost sure (a.s.) as n approaches to infinity, i.e.,

$$\lim_{n \to \infty} w(n) = w^* \quad \text{a.s.} \qquad (26)$$

with some $w^* \in W^*$ if the step size $\eta(n)$ is chosen as $\eta(n) \equiv \eta$ where

$$0 < \eta < 2. \qquad (27)$$

By virtue of (15), the property (26) yields

$$J(w) \xrightarrow[n \to \infty]{} 0 \quad \text{a.s.} \qquad (28)$$

The proof of the probabilistic convergence of $\{w(n)\}$ caused by the learning algorithm (14) with constant η satisfying (27) utilizes essentially the Doob's martingale convergence theorem [24] after establishing the fact that, under the condition (25), the random $\{V_n\}$ is the supermartingale defined as

$$E\{V_n \mid w(n-1), \ldots, w(0)\} \leq V_{n-1} \qquad (29)$$

with

$$V_n = \inf_{w^* \in W^*} \| w^* - w_n \|^2, \qquad (30)$$

and also takes into account the Borel – Cantelli lemma [24, sect. 15.3]. (In expression (29), $E\{V_n \mid \cdot\}$ denotes the conditional expectation of V_n.)

The conditions given above are only the sufficient conditions guaranteeing the local convergence of (14) with probability 1. Since the condition b) requires some computation effort for its verification whereas the condition c) cannot be verified before starting the learning algorithm, these local convergence results make of the mathematical sense.

Comparing (29) and (19), we notice that the variable V_n given by (30) is a peculiar stochastic counterpart of the Lyapunov function of (14) if $w(n) \in W_{\varepsilon}(w^*)$ will be guaranteed.

At first sight, it seems that $V(w)$ defined in (18) might be exploited as a Lyapunov function for analyzing the asymptotic behavior of (14) in a stochastic framework for any $w(0) \in \mathrm{IR}^{M(N+2)+1}$. In fact, by the definition, $V(w)$ has the property

$$V(w) = 0 \text{ if } w \in W^* \text{ and } V(w) > 0 \text{ if } w \notin W^*. \tag{31}$$

However, the requirement

$$\| \nabla V(w') - \nabla(w') \| \leq L \| w' - w'' \| \tag{32}$$

with the Lipschitz constant $L > 0$ advanced in [27] is not satisfied for any w', w'' from $\mathrm{IR}^{M(N+2)+1}$. Thus, $V(w)$ having the form (18) is indeed not admissible to study the global convergence properties of (14) based on results of [27].

In [26] it has been derived that the limit (26) will be achieved for an arbitrary initial $w(0)$ if the assumptions $1-3$ made above hold and, instead of (27), the learning rate, $\eta(n)$, is chosen as

$$0 < \eta < 2\theta / L\tau \tag{33}$$

with

$$\theta := \inf_{w \notin W^*} \frac{\| \nabla_w E\{Q(x,w)\} \|^2}{E\{Q(x,w)\}} > 0, \qquad \tau := \sup_{w \notin W^*} \frac{E\{\| \nabla_w Q(x,w) \|^2\}}{E\{Q(x,w)\}} < \infty.$$

The proof of this result establishing the conditions for the global probabilistic convergence of the learning algorithm (14) utilizes the Theorem 3′ of [27] after replacing $V(w)$ of the form (18) by

$$V(w) = E\{Q(x,w)\}. \tag{34}$$

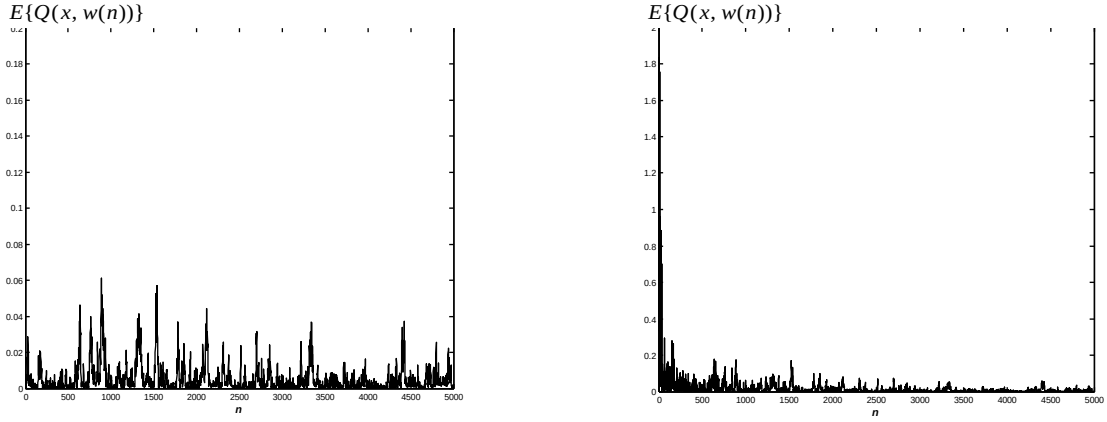
Now, let $\xi(n)$ be present. Then, the requirement (33) needs to be replaced by another requirement under which $\eta(n) \to 0$ as n tends to ∞. It can be shown that, using the same Lyapunov function as in (34), and exploiting the Theorem 3 of [27], the convergence properties (27), (29) will be ensured if $\{\eta(n)\}$ satisfies the standard requirements

$$\sum_{n=0}^{\infty} \eta(n) = \infty, \qquad \sum_{n=0}^{\infty} \eta^2(n) < \infty \tag{35}$$

arising first in [6]. (Notice that (35) are satisfied if $\eta(n) = n^{-\alpha}$ with $1/2 < \alpha < 1$.)

To illustrate the asymptotic behavior of the algorithm (14) in the presence of noise, we conducted the same simulation experiment 1 but with $\xi(n) \not\equiv 0$. Namely, $\{\xi(n)\}$ was chosen as a pseudorandom i.i.d. sequence in the range $[-0.05,\ 0.05]$. Two separate simulations were conducted. In first simulation, $\eta(n)$ was chosen as $\eta(n) \equiv 0.01$ whereas in second simulation, $\eta(n) = n^{-0.51}$ was taken.

Results of these simulation experiments are presented in Fig. 3.



a) b)

Fig. 3. Behavior of learning algorithm (14) with noise in Example 1:
a) $\eta \equiv 0.01$; b) $\eta(n) = n^{-0.51}$

It is seen that in the first case, where the learning rate remains constant, the oscillations of $E\{Q(x, w(n))\}$ are observed (Fig. 3, a). Nevertheless, this variable converges to zero as n becomes large enough; see Fig. 3, b.

Conclusion. The main contribution of this paper consisted in theoretical and experimental studying the asymptotical properties of standard online gradient algorithms applicable to the learning neural networks in the stochastic framework. Namely, sufficient conditions for the global convergence of these algorithms have been established. It was shown that adding a penalty term to the current error function is indeed not necessary to guarantee their convergence properties.

References

1. Ивахненко А.Г., Степашко В.С. Помехоустойчивость моделирования. – К.: Наук. думка, 1985. – 216 с.

2. Suykens J., Moor B.D. Nonlinear system identification using multilayer neural networks: some ideas for initial weights, number of hidden neurons and error criteria

// Proc. 12nd IFAC World Congress (Sydney, Australia, July 1993). – 1993. – vol. 3. – P. 49–52.

3. Kosmatopoulos E.S., Polycarpou M.M., Christodoulou M.A., Ioannou P.A. High-order neural network structures for identification of dynamical systems // IEEE Trans. on Neural Networks. – 1995. – vol. 6. – P. 422–431.

4. Levin A.U., Narendra K. S., Recursive identification using feedforward neural networks // Int. J. Contr. – 1995. – vol. 61. – P. 533–547.

5. Tsypkin Ya.Z., Mason J.D., Avedyan E.D., Warwick K., Levin I. K. Neural networks for identification of nonlinear systems under random piecewise polynomial disturbances // IEEE Trans. on Neural Networks. – 1999. – vol. 10. – P. 303–311.

6. Tsypkin Ya. Z. Adaptation and learning in automatic systems. – New-York: Academic Press. – 1971. – 291 p.

7. White H. Some asymptotic results for learning in single hidden-layer neural network models // J. Amer. Statist. Assoc. – 1987. – vol. 84. – P. 117–134.

8. Behera L., Kumar S., Patnaik A. On adaptive learning rate that guarantees convergence in feedforward networks // IEEE Trans. on Neural Networks. – 2006. – vol. 17. – P. 1116–1125.

9. Kuan C. M., Hornik K. Convergence of learning algorithms with constant learning rates // Ibid. – 1991. – vol. 2. – P. 484 – 489.

10. Luo Z. On the convergence of the LMS algorithm with adaptive learning rate for linear feedforward networks // Neural Comput. – 1991. – vol. 3. – P. 226–245.

11. Finnoff W. Diffusion approximations for the constant learning rate backpropagation algorithm and resistance to local minima // Ibid. – 1994. – 6. – P. 285– 295.

12. Gaivoronski A.A. Convergence properties of backpropagation for neural nets via theory of stochastic gradient methods // Optim. Methods Software. – 1994. – 4. – P. 117–134.

13. Fine T.L., Mukherjee S. Parameter convergence and learning curves for neural networks // Neural Comput. – 1999. – 11. – P. 749–769.

14. Tadic V., Stankovic S. Learning in neural networks by normalized stochastic gradient algorithm: Local convergence // Proc. 5th Seminar Neural Netw. Appl. Electr. Eng. (Yugoslavia,Sept. 2000). – 2000. – P. 11–17.

15. Zhang H., Wu W., Liu F., Yao M. Boundedness and convergence of online gradient method with penalty for feedforward neural networks // IEEE Trans. on Neural Networks. – 2009. – vol. 20. – P. 1050–1054.

16. Mangasarian O.L., Solodov M.V. Serial and parallel backpropagation convergence via nonmonotone perturbed minimization // Optim. Methods Software. –1994. – P. 103–106.

17. Wu W., Feng G., Li X. Training multilayer perceptrons via minimization of ridge functions // Advances in Comput. Mathematics. – 2002. – vol. 17. – P. 331–347.

18. Zhang N., Wu W., Zheng G. Convergence of gradient method with momentum for two-layer feedforward neural networks // IEEE Trans. on Neural Networks. – 2006. – vol. 17. – P. 522–525.

19. Wu W., Feng G., Li X., Xu Y. Deterministic convergence of an online gradient method for BP neural networks // Ibid. –2005. – vol. 16. – P. 1–9.

20. Xu Z.B., Zhang R., Jing W.F. When does online BP training converge? // Ibid. – 2009. – vol. 20. – P. 1529–1539.

21. Shao H., Wu W., Liu L. Convergence and monotonicity of an online gradient method with penalty for neural networks // WSEAS Trans. Math. – 2007. – vol. 6. – P. 469–476.

22. Ellacott S.W. The numerical analysis approach // Mathematical Approaches to Neural Networks (J.G. Taylor, ed; B.V.: Elsevier Science Publisher). – 1993. – P. 103–137.

23. Skantze F.P., Kojic A., Loh A.P., Annaswamy A.M. Adaptive estimation of discrete time systems with nonlinear parameterization // Automatica. – 2000. – vol. 36. – P. 1879–1887.

24. Loeve M. Probability theory. – New-York: Springer-Verlag. – 1963. – 425 p.

25. Zhiteckii L.S., Azarskov V.N., Nikolaienko S.A. Convergence of learning algorithms in neural networks for adaptive identification of nonlinearly parameterized systems // in Proc. 16th IFAC Symposium on System Identification (Brussels, Belgium). – 2012. – P. 1593–1598.

26. Azarskov V.N., Kucherov D.P, Nikolaienko S.A., Zhiteckii L.S. Asymptotic behaviour of gradient learning algorithms in neural network models for the identification of nonlinear systems // American Journal of Neural Networks and Applications. – 2015. – No 1(1). – P. 1–10.

27. Polyak B.T. Convergence and convergence rate of iterative stochastic algorithms, I: General case // Autom. Remote Control. – 1976. – vol. 12. – P. 1858–1868.