American Journal of Neural Networks and Applications
2015; 1(1): 1-10
Published online July 27, 2015 (http://www.sciencepublishinggroup.com/j/ajnna)
doi: 10.11648/j.ajnna.20150101.11

Asymptotic Behaviour of Gradient Learning Algorithms in Neural Network Models for the Identification of Nonlinear Systems

Valerii N. Azarskov[1], Dmytro P. Kucherov[1], Sergii A. Nikolaienko[2], Leonid S. Zhiteckii[2]

[1]Faculty of Computer Science, National Aviation University, Kiev, Ukraine
[2]Cybernetics Centre, Dept. of Automated Data Processing Systems, Kiev, Ukraine

Email address

azarskov@nau.edu.ua (V. N. Azarskov), d_kucherov@ukr.net (D. P. Kucherov), s_nikolaenko@ukr.net (S. A. Nikolaienko), leonid_zhiteckii@i.ua (L. S. Zhiteckii)

To cite this article:

Valerii N. Azarskov, Dmytro P. Kucherov, Sergii A. Nikolaienko, Leonid S. Zhiteckii. Asymptotic Behaviour of Gradient Learning Algorithms in Neural Network Models for the Identification of Nonlinear Systems. *American Journal of Neural Networks and Applications*. Vol. 1, No. 1, 2015, pp. 1-10. doi: 10.11648/j.ajnna.20150101.11

Abstract: This paper deals with studying the asymptotical properties of multilayer neural networks models used for the adaptive identification of wide class of nonlinearly parameterized systems in stochastic environment. To adjust the neural network's weights, the standard online gradient type learning algorithms are employed. The learning set is assumed to be infinite but bounded. The Lyapunov-like tool is utilized to analyze the ultimate behaviour of learning processes in the presence of stochastic input variables. New sufficient conditions guaranteeing the global convergence of these algorithms in the stochastic frameworks are derived. The main their feature is that they need no a penalty term to achieve the boundedness of weight sequence. To demonstrate asymptotic behaviour of the learning algorithms and support the theoretical studies, some simulation examples are also given.

Keywords: Neural Network, Nonlinear Model, Gradient Learning Algorithm, Stochastic Environment, Convergence

1. Introduction

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 [1–4]. This has been motivated by the theoretical works of several researches including, in particular, Cybenko and Funahashi [5, 6] who proved that, even with one hidden layer, neural network can uniformly approximate any continuous mapping over a compact domain, provided that the network has sufficient number of neurons with corresponding weights.

Several learning methods for updating the weights of neural networks have been advanced in literature. Most of these methods rely on the gradient concept [4, 7]. 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 [2, 8].

The asymptotic behaviour of online adaptive gradient algorithms for the network learning has been studied by many authors. In particular, White [9] investigated the convergence of the learning process for the so-called feedforward network models with single hidden layer by using the stochastic approximation theory. The convergence results have been derived in [10–16] 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 [17–22]. In contrast to the stochastic approach, several of these results allow to employ a constant learning rate [19, 23]. 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 this difficulties in non-stochastic case, the assumption that similar nonlinear functions need to be convex (concave) is introduced in [24]. However, such an assumption is not appropriate for neural network's description of nonlinearity.

A popular approach to analyze the asymptotic behaviour of online gradient algorithms in stochastic case is based on Martingale convergence theory [25]. This approach has been exploited in [26, 27] to derive some local convergence in stochastic framework for standard online gradient algorithms with the constant learning rate.

This paper is an extension of [26, 27]. 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 is ensured. The key idea in deriving these convergence results is based on the use of the Lyapunov methodology [28].

2. The Description of System

Let

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

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 $(n = 1, 2, \ldots)$, 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 [7, subsect. 5.15].) Without loss of generality, one supposes that the nonlinearity

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

is the continuous and smooth function on a bounded but infinite set $X \subset \mathrm{IR}^N$ (diam $X < \infty$).

To approximate (2) 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 jth neuron of this layer at the time instant n are the components of $x(n)$. Its output signal at the nth time instant is specified as

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

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

$$\sigma(s) = \frac{1}{1 + \exp(-s)} \tag{4}$$

or

$$\sigma(s) = \tanh(s). \tag{5}$$

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

$$y^{(2)}(n) = \sum_{j=1}^{M} w_j^{(2)} y_j^{(1)}(n) + b^{(2)}, \tag{6}$$

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 (4) and (5) are nonlinear, it follows from (3), (6) that $y^{(2)}(n)$ is the nonlinear function depending on $x(n-1)$ and also on the $(M(N+2)+1)$ -dimensional parameter vector

$$\begin{aligned} w = [&w_{11}^{(1)}, \ldots, w_{N1}^{(1)}, b_1^{(1)}, \ldots, w_{1M}^{(1)}, \ldots, w_{NM}^{(1)}, b_M^{(1)} \vdots \\ &w_1^{(2)}, \ldots, w_M^{(2)}, b^{(2)}]^T. \end{aligned} \tag{7}$$

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

$$y^{(2)}(n) = \mathrm{NN}(x(n), w) \tag{8}$$

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 (1), rewrite (8) as follows:

$$y_{\mathrm{mod}}(n) = \mathrm{NN}(x(n), w). \tag{9}$$

Now, define the variable

$$e = F(x) - \mathrm{NN}(x, w) \tag{10}$$

representing the discrepancy between the nonlinearity (2) and its neural network's model for a fixed w. Due to (1), it yields the current model error

$$e(n) = y(n) - \mathrm{NN}(x(n), w) \tag{11}$$

which can be measured at the nth time instant. Further, introduce the usual quadratic loss function

$$Q(x, w) = [F(x) - \mathrm{NN}(x, w)]^2. \tag{12}$$

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

$$w(n) = w(n-1) - \eta(n) \nabla_w Q(x(n), w(n-1)) \tag{13}$$

taken, for example, from [4, 7] is utilized. In this algorithm,

$\nabla_w Q(x(n), w(n-1))$ denotes the gradient of $Q(x, w)$ with respect to w at $w = w(n-1)$ for given $x = x(n)$, and $\eta(n)$ is the learning rate (step size) of (13). Thus, (3), (6), (8) and (13) together with (9) and (12) 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.

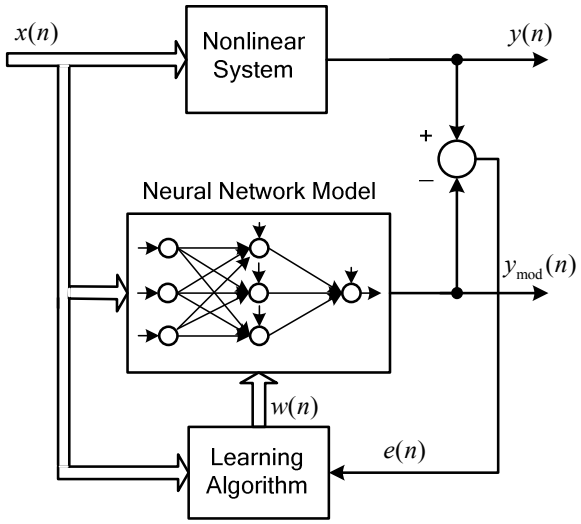


Figure 1. *Configuration of online learning system.*

3. Problem Formulation

Let $\{x(n)\}$ be 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.

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

Now, introduce the performance index

$$J(w) = E\{Q(x, w)\} \tag{14}$$

which evaluates the quality of learning process with $Q(x, w)$ given in (12). 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.

The aim of this paper consists in studying the asymptotic properties of the learning procedure (13). More certainty, the following problem is stated. It is required to derive the conditions under which $\{w(n)\}$ caused by this recursive algorithm will converge in the sense that

$$J(w(n)) \to \inf_w J(w) \text{ as } n \to \infty \tag{15}$$

with probability 1 (almost sure), where $J(w(n))$ is determined by (14) for $w = w(n)$.

4. Preliminaries

Let w^* be a vector minimizing $J(w)$, i.e.,

$$w^* = \arg\inf_w J(w). \tag{16}$$

Consider, first, the case when $F(x)$ can exactly be approximated by a neural network representation for all $x \in X$ implying

$$F(x) \equiv \mathrm{NN}(x, w^*). \tag{17}$$

In this case called in [4, p. 304] as the ideal case, one has $J(w^*) = 0$ (by virtue of (12), (14)).

It turned out that, at least, in the ideal case, the set W^*, containing these w^*s becomes not one-point [26, 27]. To show it, put $N = 1$, $M = 1$. Due to (7), this implies $w^* \in \mathrm{IR}^4$. Let $w^* = [w_1^*, w_2^*, w_3^*, w_4^*]^T$ be a vector satisfying (17). Then, (3) and (6) together with (4) give that another $w^* = [-w_1^*, -w_2^*, -w_3^*, w_3^* + w_4^*]^T$ will also satisfy (17).

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.,

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

where V_∞ is a random value (in general), meaning that the algorithm (13) converges. On the other hand, the fact that $\{V_n\}$ is monotonically 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 (17) is satisfied. Hence,

$$w(n) \xrightarrow[n\to\infty]{} w_\infty \tag{21}$$

with $w_\infty = w^*$ is not guaranteed without additional assumptions on $\{x(n)\}$. Moreover, the limit (20) may not exist if $\{x(n)\}$ is an arbitrary non-stochastic sequence leading to the violation of (19) [26]. Nevertheless, if the asymptotic property (21) takes place, then $\{w(n)\}$ converges to some $w_\infty \in \lim\inf W_n$ in sense of (21) where

$$\lim\inf W_n := \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} W_k \tag{22}$$

denotes the so-called limit set introduced in [27, sect. 1.3] in which

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

Comment 1. Note that the limit set, $\lim\inf W_n$, given by (22) represents a nonlinear manifold on $\mathrm{IR}^{M(N+2)+1}$ whose dimension satisfies $0 \le \dim \lim\inf W_n \le M(N+2)$.

It can be understood that the algorithm (13) "attempts" to solve the infinite set of the equations

$$y(n) - \mathrm{NN}(x(n-1), w) = 0, \quad n = 1, 2, \ldots \tag{23}$$

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 (23), which is determined as the limit set (22) 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 neither for non-stochastic nor for stochastic $\{x(n)\}$s.

In [26, 27], it has been established that if initial $w(0)$ is chosen at an ε-neighbourhood, $U_\varepsilon(w^{*(i)}) := \{w : \| w^{*(i)} - w \| < \varepsilon\}$, of some $w^{*(i)} \in W^*$ giving $V_0 = \| w^{*(i)} - w(0) \|^2$ (according to (18)), then $\{V_n\}$ behaves as the so-called positive supermartingale [25, 29] if $\{x(n)\}$ is a stochastic sequence. This implies that the conditional expectation satisfies

$$E\{V_n \mid V_{n-1}, \ldots, V_0\} \le V_{n-1}. \tag{24}$$

Notice that (24) represents a stochastic counterpart of (19).) By virtue of the well-known Doob's theorem [25], the property (24) yields

$$\lim_{n\to\infty} V_n = V_\infty \ \text{ a.s.} \tag{25}$$

similar to (20) for non-stochastic case. However, if $w(0)$ lies far enough from W^*, then the condition (24) may not be satisfied. In this case, instead of (24), other condition

$$E\{V_n \mid V_{n-1}, \ldots, V_0\} \le V_n + \chi_n, \quad \chi_n \ge 0 \tag{26}$$

which is more strong may take place.

The asymptotic behaviour of the gradient algorithm (13) for an arbitrary $w(0)$ is now derived in the following theorem.

Theorem 1. Let (17) hold and

$$\sum_{n=0}^{\infty} P\{\chi_n > 0\} < \infty \ \text{ a.s.} \tag{27}$$

Then $\{V_n\}$ will converge with probability 1 (a.s.) for any $w(0)$ provided that the condition (26) is satisfied.

Proof. Using the classical Borel-Cantelli lemma [25, Sect. 15.3], from (27) it can conclude that there exists a finite number $n^* < \infty$ such that $\chi n = 0$ for all $n \ge n^*$. Since $\{\chi n\}$ is nonnegative, this gives

$$\sum_{n=0}^{\infty} \chi_n < \infty \ \text{ a.s.} \tag{28}$$

meaning that the all conditions of the modified positive supermartingale result established in Corollary D.5.1 of [29, p.501] are satisfied. By this result, the validity of (25) follows.

Comment 2. Of course, the convergence conditions (27) given in Theorem 1 (or (28)) make only the mathematical sense because they cannot beforehand be verified. Nevertheless, this conditions are somewhat useful for understanding the asymptotic behaviour of the learning algorithm (13) in the stochastic environment.

At first sight, it seems that the variable V(w) given by (18) might be exploited as a Lyapunov function for analyzing the asymptotic behaviour of (13) in a stochastic framework. In fact, by the definition, V(w) has the property

$$V(w) = 0 \ \text{ if } w \in W^* \text{ and } V(w) > 0 \ \text{ if } w \notin W^* \tag{(29)}$$

Meanwhile, the partial derivatives of V(w) with respect to the components of W are not continued at all $w \in \mathrm{IR}^{M(N+2)+1}$. To demonstrate this feature, let $W^* = \{w^{*(1)},\ w^{*(2)}\}$. Then $V(w)$ becomes

$$V(w) = \min\{V^{(1)}(w),\ V^{(2)}(w)\},$$

where $V^{(i)}(w) = \| w^{*(i)} - w \|^2 \quad (i = 1, 2)$. In this case, the components of the gradient ∇V(w) which represent these partial derivatives are discontinuous at ws belonging to the boundary between the domains $W^{(1)} = \{w:\ V^{(1)} < V^{(2)}\}$ and $W^{(2)} = \{w:\ V^{(2)} < V^{(1)}\}$ (see Fig. 2). Thereby, the requirement

$$\| \nabla V(w') - \nabla(w') \| \le L \| w' - w'' \| \tag{30}$$

with the Lipschitz constant $L > 0$ advanced in [28] 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 (13) based on results of [28].

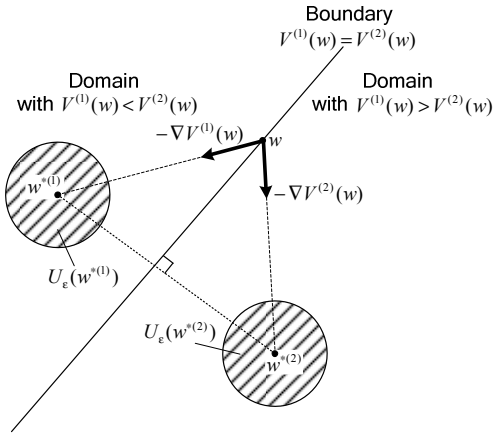


Figure 2. *I llustration of the two-layers networks properties with M=1, N=1.*

5. Observations

To demonstrate some asymptotic properties of (13) pointed out in the section above, several 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 (3), (4), (6) and (8) with $w^{*(1)}$ and $w^{*(2)}$ given as: $w^{*(1)}=[7.15,\ 1.65,\ 3.45,\ 0.3]^T$ and $w^{*(2)}=[-7.15,\ -1.65,\ -3.45,\ 3.75]^T$.

In all of the experiments, η was taken as η=0.01.

Fig. 3 illustrated the results of the first simulation example, where $\{x(n)\}$ was chosen as a non-stochastic sequence. It can be observed that in this example V_n defined by (18) with $w=w(n)$ no limit implying that the learning algorithm (13) is not converge.

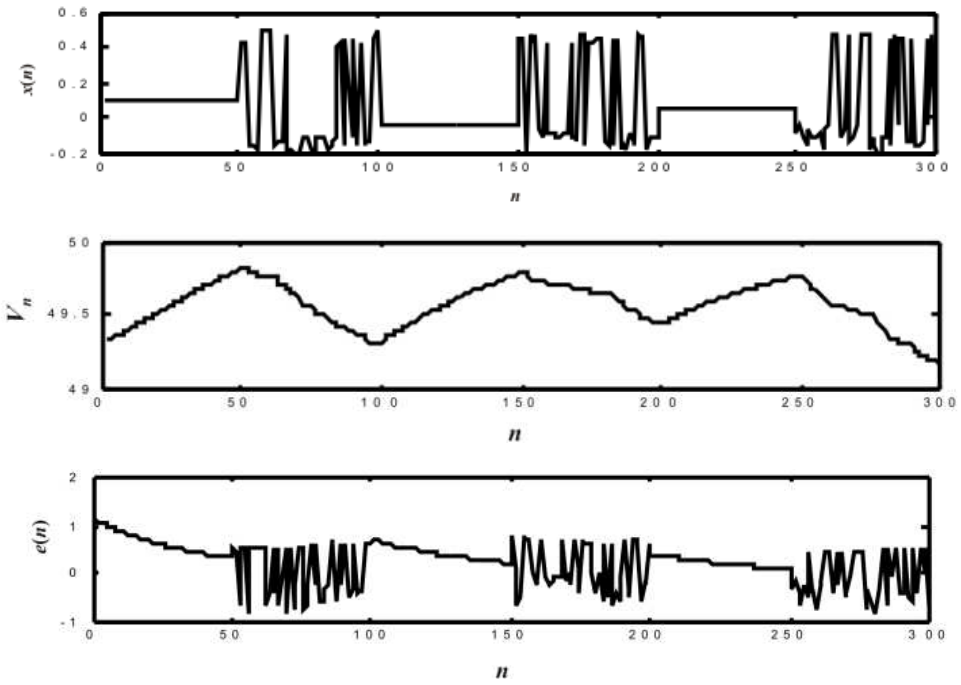


Figure 3. *Behaviour of gradient learning algorithm (13) in Example 1.*

In other experiments, $\{x(n)\}$ was generated as sequence of independent identically distributed (i.i.d.) pseudo random numbers on $X=[-1.0,\ 1.0]$ (the stochastic cases). Namely, in the second example, the initial $w(0)$ was taken closely to $w^*(1)$. In this case, the first difference $\Delta V_n=V_n-V_{n-1}$ of V_n defined by (18) changed its sign (see Fig. 4). Nevertheless, V_n tends to zero and $e(n)\to 0$ as n increases as shown in Fig. 4. This observation supports the convergence property of (13) showing that $\{V_n\}$ is here the supermartingale.

Simulation results of third and fourth experiments are presented in Fig. 5 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. 5, left. Further, $\{V_n^{(1)}\}$ became decreasing. Such a behaviour of these sequence leaded to appearing the feature that $V_n^{(1)}<V_n^{(2)}$ for all sufficiently large n.

In the fourth example, the initial w(0) was chosen to be

close to that in the third example. One can observe that in this case, $V_n \equiv V_n^{(1)}$ (see Fig. 5, right).

It turned out that in third and fourth simulation examples, the condition (24) is not satisfied whereas the learning algorithm (13) remains indeed convergent. In these examples, instead of (24), the condition (26) takes place. This fact is demonstrated in Fig. 6.

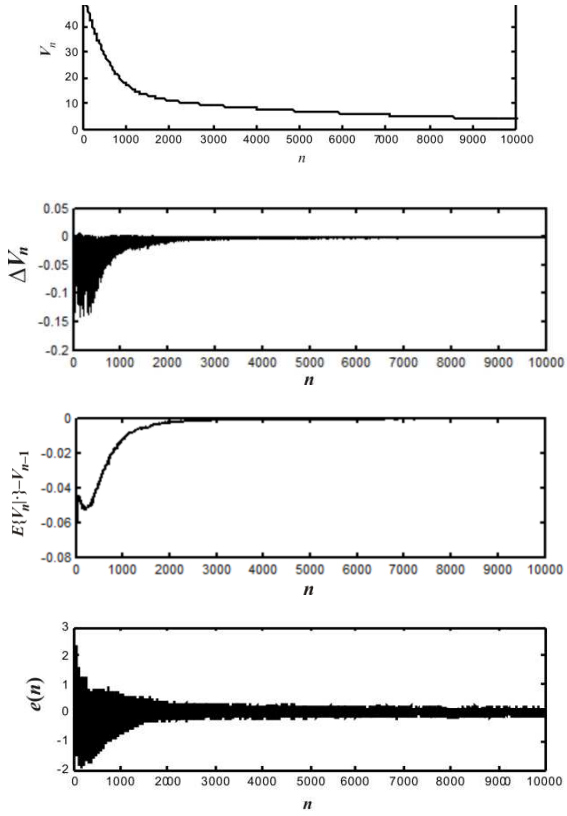


Figure 4. *Behaviour of gradient learning algorithm (13) in Example 2.*

6. Main Results

The global stochastic convergence analysis of the gradient learning algorithm (13) is based on employing the fundamental convergence conditions established in the Key Technical Lemma which is the slightly reformulated Theorem 3 of [28].

Key Technical Lemma. Let V(w) be a function satisfying (29) and (30). Define the scalar variable

$$H(w) = \nabla_w V(w)^T \nabla_w E\{Q(x,\ w)\} \tag{31}$$

and denote

$$H_n(w) := \nabla_w V(w(n))^T \nabla_w E\{Q(x,\ w(n))\}.$$

Suppose:

(i) $H_n(w) \geq \theta_n V(w(n-1)),\quad \theta_n > 0,$

(ii) $E\{\| \nabla_w Q(x,\ w(n)) \|^2\} \leq \tau_n V(w(n)),\quad \tau_n \geq 0.$

Introduce the additional variable

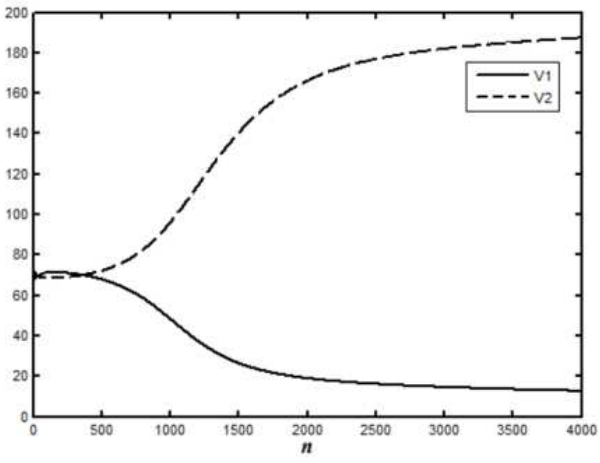
$$\nu_n = \eta(n)(\theta_n - L\eta(n)\tau_n/2). \tag{32}$$

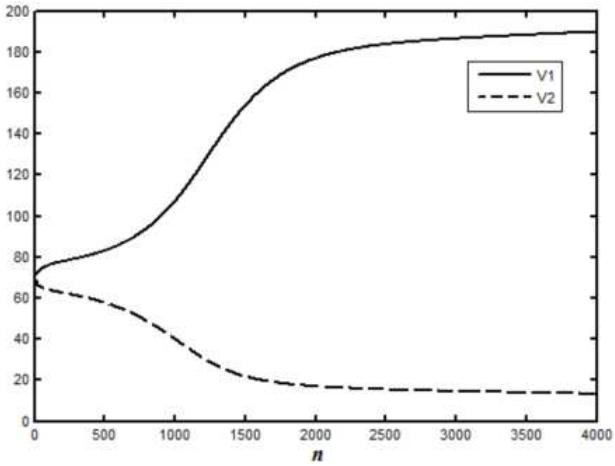
Then the algorithm (13) yields $\lim_{n\to\infty} V_n = 0$ a.s. provided that $E\{w(0)\} < \infty$ and

$$0 \leq \nu_n \leq 1, \tag{33}$$

$$\sum_{n=0}^{\infty} \nu_n = \infty, \tag{34}$$

i.e., the limit (25) will be achieved for $V_\infty = 0$.





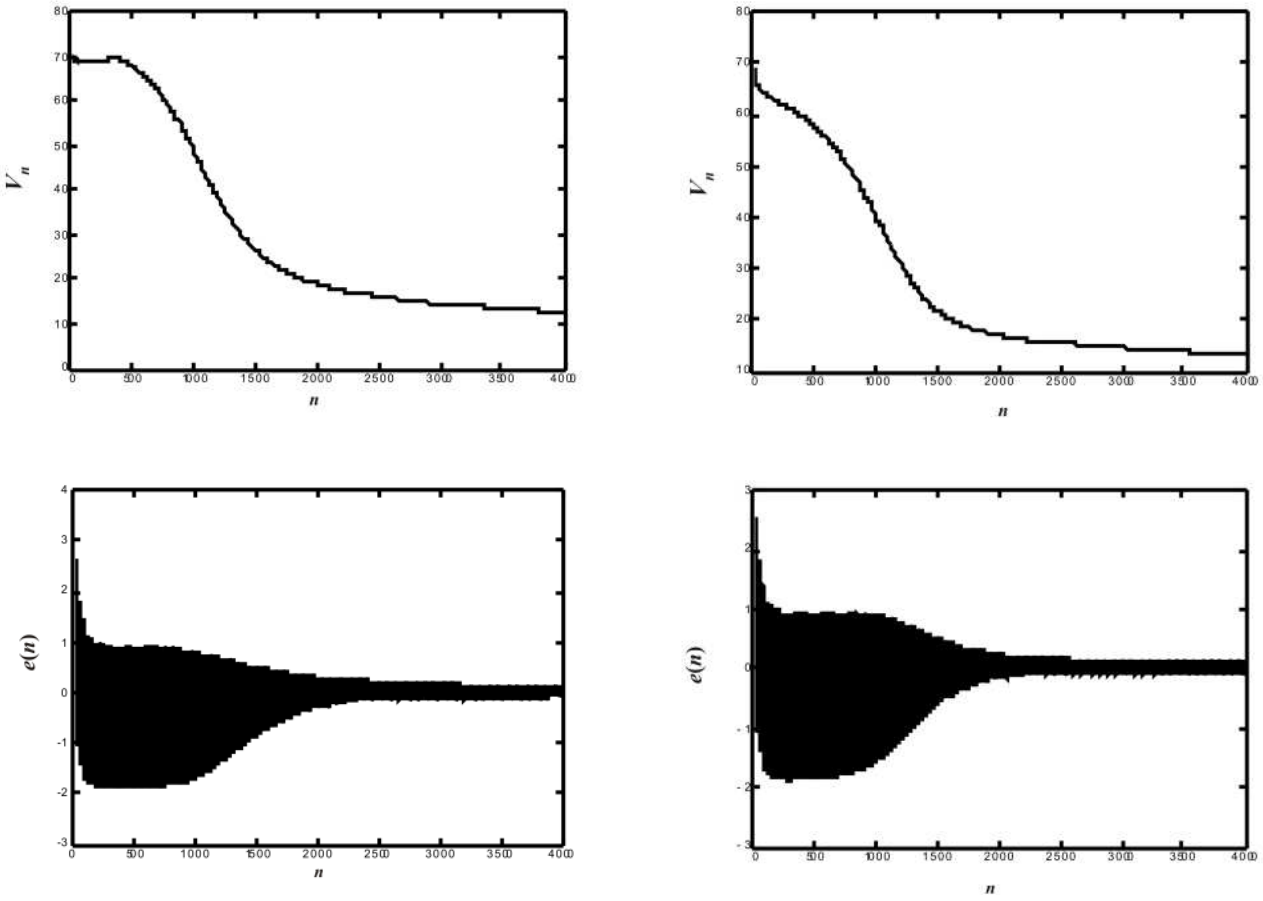


Figure 5. *Behaviour of gradient learning algorithm (13) in Examples 3 (left) and 4 (right).*

Related results followed from the Theorem 3′ of (28) are:

Corollary. Under the conditions of the Key Technical Lemma, if $\theta_n \equiv \theta = \text{const}$ and $\tau_n \equiv \tau = \text{const}$, and $\eta(n) \equiv \eta = \text{const}$, then $V_n \xrightarrow[n\to\infty]{} 0$ a.s. provided that

$$0 < \eta \le 2(\theta - \varepsilon)/L\tau \quad (0 < \varepsilon < \theta) \tag{35}$$

is satisfied.

Now, one is able to present the first convergence result summarized in the theorem below.

Theorem 2. Suppose the assumption (17) holds. Then the gradient algorithm (13) with a constant learning rate, $\eta(n) \equiv \eta$, will converge with probability 1 (in the sense that $V_n \xrightarrow[n\to\infty]{} 0$ a.s.) and

$$\lim_{n\to\infty} e(n) = 0 \quad \text{a.s.} \tag{36}$$

for any initial $w(0)$ chosen randomly so that $E\{Q(x, w(0))\} < \infty$ if the conditions (35) with θ and τ specified by

$$\theta := \inf_{w \notin W^*} \frac{\| \nabla_w E\{Q(x,w)\} \|^2}{E\{Q(x,w)\}}, \tag{37}$$

$$\tau := \sup_{w \notin W^*} \frac{E\{\| \nabla_w Q(x,w) \|^2\}}{E\{Q(x,w)\}} \tag{38}$$

are satisfied.

Proof. Set

$$V(w) = E\{Q(x,w)\}. \tag{39}$$

Then condition (29) and (30) can be shown to be valid. This indicates that $V(w)$ of the form (39) may be taken as the Lyapunov function. By virtue of (31) such a choice of $V(w)$ gives $H(w) = \| \nabla_w E\{Q(x,w)\} \|^2$. Putting $\theta_n \equiv \theta$ and $\tau_n \equiv \tau$ with θ and τ determined by (37) and (38), respectively, one can conclude that the conditions (i), (ii) of the Key Technical Lemma are satisfied. Applying its Corollary it proves that $\lim_{n\to\infty} V_n = 0$ with probability 1.

Due to the definition (39) of $V(w)$ together with the assumption (17), result (36) follows.

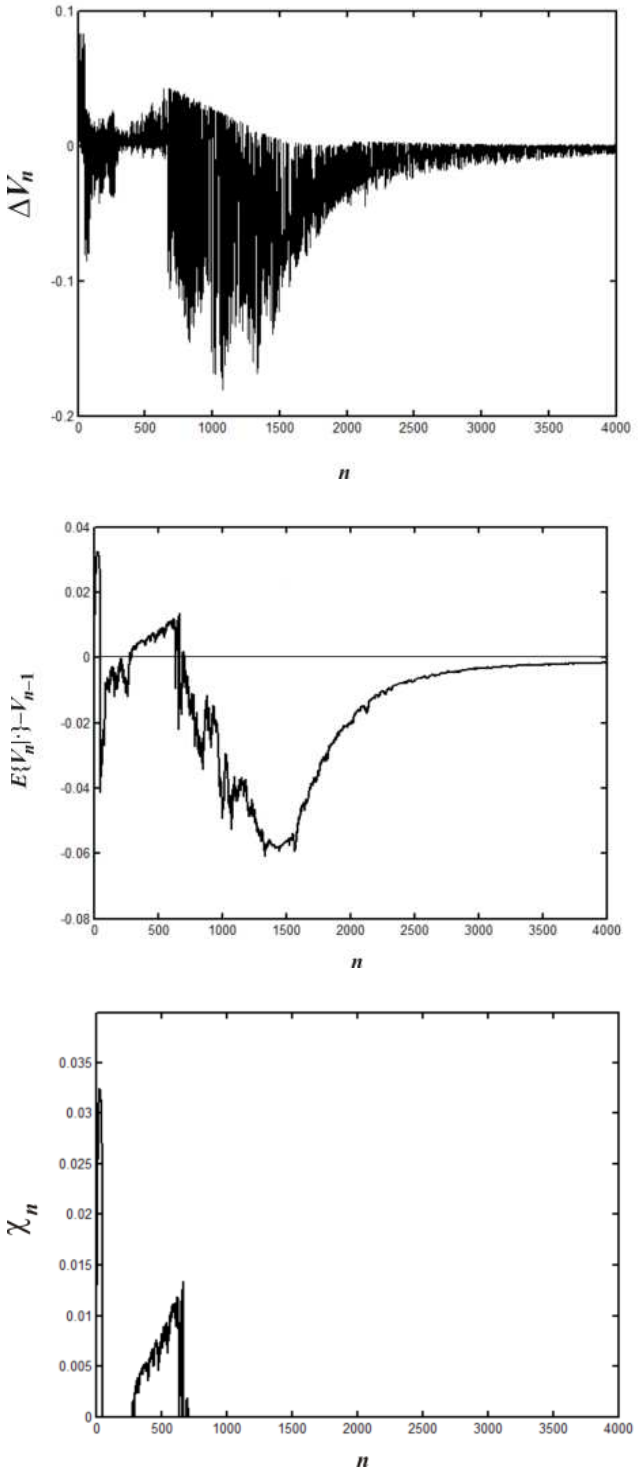


Figure 6. *Variables* ΔV_n, $E\{V_n|\cdot\}-V_{n-1}$ *and* χ_n *in Example 3.*

Now, consider general case, where $F(x)$ cannot exactly be approximated by $\mathrm{NN}(x, w)$ (as in (17)). Obviously, in this case, $\inf_w Q(x, w^*) \not\equiv 0,$ and the choice of a constant learning rate, $\eta(n) = \eta,$ is not appropriate [4].

The convergence results are here established in the follow theorem.

Theorem 3. Subject to the conditions

$$\text{(a)} \sum_{n=0}^{\infty} \eta(n) = \infty, \quad \text{(b)} \sum_{n=0}^{\infty} \eta^2(n) < \infty, \tag{40}$$

the gradient algorithm (13) yields

$$\lim_{n\to\infty} E\{Q(x, w(n))\} = \inf_w E\{Q(x, w))\} \quad \text{a.s.}$$

provided that $\theta > 0$ with θ determined by (37).

Proof. Setting

$$V_n = E\{Q(x, w(n))\} - \inf_w E\{Q(x, w)\}$$

it can show that the requirements (29) and (30) will be satisfied: $V(w^*) = 0,$ and $V(w) > 0$ for $w \neq w^*$. Since $E\{\| \nabla_w Q(x, w) \|\}^2 > 0$ for $w = w^*,$ it follows that condition (ii) of the Key Technical Lemma assumes $\tau_n \to \infty$ as $w(n) \to w^*$.

Suppose (ii) is not satisfied. Then, there is a finite $\overline{\tau}$ such that

$$E\{\| \nabla_w Q(x, w) \|\}^2 \le \tau_n V(w(n))$$

With

$$\tau_n \le \overline{\tau} < \infty. \tag{41}$$

Since τn is assumed to be finite, there exists a finite n_0 such that requirement (33) will be satisfied for all sufficiently large $n \ge n_0$ provided that (i) takes place with $\theta_n \ge \theta > 0$ and $E\{w(n_0)\} < \infty$ and the condition (b) of (40) is satisfied (due to the fact that (b) means $\eta(n) \to 0$ as $n \to \infty$).

Further, if the assumption $\tau_n \le \overline{\tau} < \infty$ holds then the series

$$\sum_{n=n_0}^{\infty} \eta_n \theta_n \quad \text{with} \quad \theta_n \ge \theta > 0$$

diverges whereas the series

$$-\sum_{n=n_0}^{\infty} L\eta(n)\tau_n / 2$$

converges (because of the validity of (a)). This gives that (34) takes also place.

Since $\theta > 0,$ all the conditions of Key Technical Lemma are satisfied for $n \ge n_0$. By this Lemma, $\lim_{n\to\infty} V_n = 0$ a.s. Therefore, $\tau_n \xrightarrow[n\to\infty]{} \infty$ with probability 1. But this contradicts the assumption that $\tau_n \le \overline{\tau} < \infty$ (see (41)). Hence, this assumption is false. This fact proves the validity of result given in theorem.

Remark. Setting

$$\theta_n := \| \nabla_w E\{Q(x, w(n))\} \|^2 / E\{Q(x, w(n))\},$$

$$\tau_n := E\{\| \nabla_w Q(x, w(n)) \|^2\} / E\{Q(x, w(n))\}$$

it can be concluded that, under the condition of the Theorem 3, the following features are observed: $\theta_n > \theta > 0,$ $\tau_n < \tau < \infty$

for all n.

Fig. 7 in which $E\{Q(x,w(n))\}$ equal to V_n demonstrates these features.

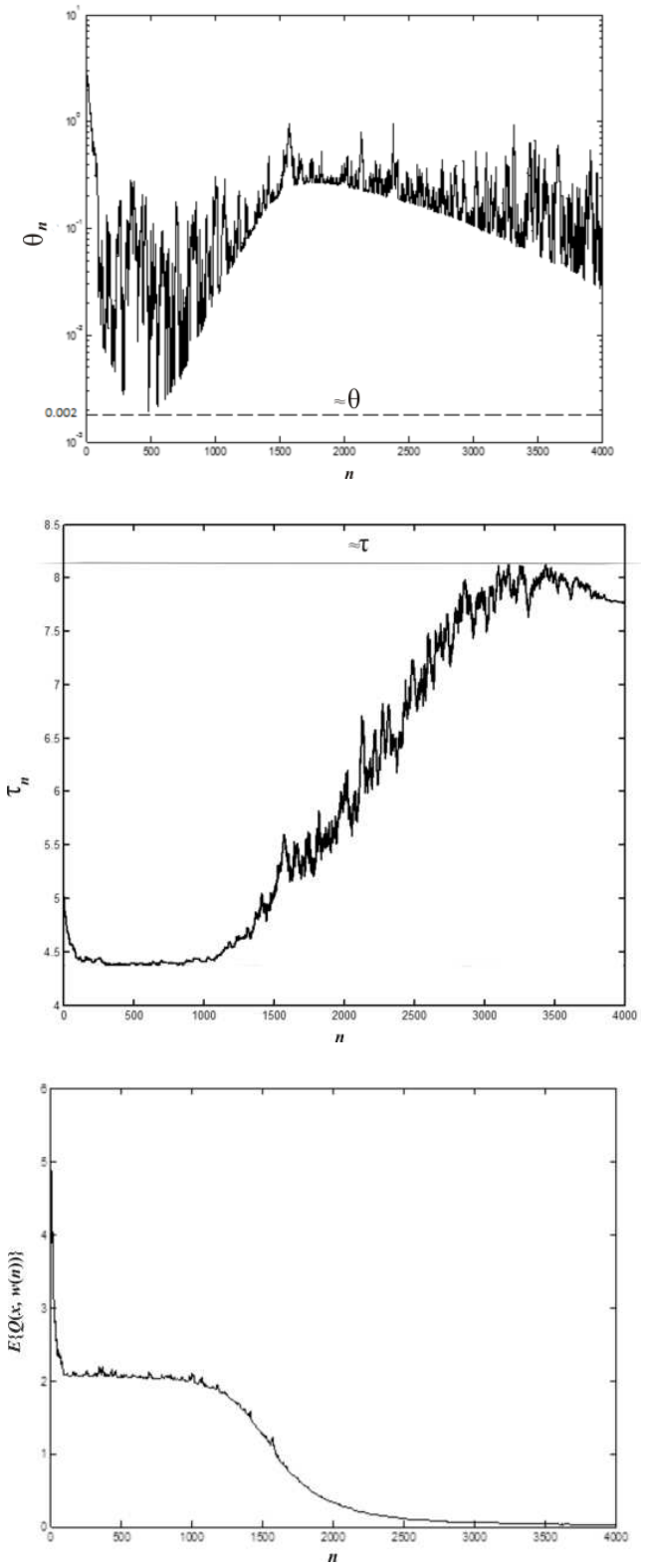


Figure 7. *Variables* θ_n, τ_n *and* $E\{Q(x,w(n))\}$ *in Example 3.*

As it is seen, $\{\theta_n\}$ is bounded away from zero whereas $\{\tau_n\}$ is upper bounded. It gives $E\{Q(x,w(n))\} \xrightarrow[n\to\infty]{} 0$ as shown in Fig. 7 below.

Comment 3. The conditions established in the theorem 2 and 3 are sufficient to guarantee the global convergence of (13) (for any $w(0)$) with probability 1 both in ideal and non-ideal cases. Under these conditions, the requirement (15) in which

$$J(w(n)) \equiv E\{Q(x,w(n))\}$$

will obviously be satisfied (final result). Again, the essential feature of this result in the fact that these convergence properties can be achieved without adding penalty term to $Q(x,w(n))$, as in [16].

Of course, the calculation of θ and τ for choosing the suitable constant learning rate, η, according to (37), (38) seems to be hard. Meanwhile, η may be replaced by the time-varying $\eta(n)$ satisfying the requirements (35) if necessary. Note that they are usual in the stochastic learning theory [7].

7. 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, new 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. Further analysis will provide a study of the asymptotic behaviour of online gradient learning algorithms in the presence of noise whose importance was pointed out in [4].

Acknowledgment

The authors would like to thank Prof. S.V. Kovalevskyy for his invitation to submit this work.

References

[1] J. Suykens, and B. D. Moor, "Nonlinear system identification using multilayer neural networks: some ideas for initial weights, number of hidden neurons and error criteria," in Proc. 12nd IFAC World Congress, vol. 3. Sydney, Australia, July 1993, pp. 49–52.

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

[3] A. U. Levin, and K. S. Narendra, "Recursive identification using feedforward neural networks," Int. J. Contr vol. 61, pp. 533–547, 1995.

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

[5] G. Cybenko, "Approximation by superpositions of a sigmoidal functions," Math. Control, Signals, Syst., vol. 2, pp. 303–313, 1989.

[6] K. Funahashi, "On the approximate realization of continuous mappings by neural networks," Neural Networks, vol. 2, pp. 182–192, 1989.

[7] Ya. Z. Tsypkin, Adaptation and Learning in Automatic Systems, New-York: Academic Press, 1971.

[8] 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, 2006.

[9] H. White, "Some asymptotic results for learning in single hidden-layer neural network models," J. Amer. Statist. Assoc., vol. 84, pp. 117–134, 1987.

[10] C. M. Kuan, and K. Hornik, "Convergence of learning algorithms with constant learning rates," IEEE Trans. on Neural Networks, vol. 2, pp. 484 – 489, 1991.

[11] Z. Luo, "On the convergence of the LMS algorithm with adaptive learning rate for linear feedforward networks," Neural Comput., vol. 3, pp. 226–245, 1991.

[12] W. Finnoff, "Diffusion approximations for the constant learning rate backpropagation algorithm and resistance to local minima," Neural Comput., 6, pp. 285– 295, 1994.

[13] A. A. Gaivoronski, "Convergence properties of backpropagation for neural nets via theory of stochastic gradient methods," Optim. Methods Software 4, pp. 117–134, 1994.

[14] T. L. Fine, and S. Mukherjee, "Parameter convergence and learning curves for neural networks," Neural Comput. 11, pp. 749–769, 1999.

[15] 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., pp. 11–17, (Yugoslavia,Sept. 2000).

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

[17] O. L. Mangasarian, and M. V.Solodov, "Serial and parallel backpropagation convergence via nonmonotone perturbed minimization," Optim. Methods Software, pp. 103–106, 1994.

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

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

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

[21] Z. B. Xu, R. Zhang, and W.F. Jing, "When does online BP training converge?" IEEE Trans. on Neural Networks, vol. 20, pp. 1529–1539, 2009.

[22] 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.

[23] S. W. Ellacott, "The numerical analysis approach," Mathematical Approaches to Neural Networks (J.G. Taylor, ed; B.V.: Elsevier Science Publisher), pp. 103–137, 1993.

[24] F. P.Skantze, A. Kojic, A. P. Loh, and A. M. Annaswamy, "Adaptive estimation of discrete time systems with nonlinear parameterization," Automatica, vol. 36, pp. 1879–1887, 2000.

[25] M. Loeve, Probability Theory New-York: Springer-Verlag, 1963.

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

[27] V. N. Azarskov, L. S. Zhiteckii, and S. A. Nikolaienko, "Sequential learning processes in neural networks applied as models of nonlinear systems," Electronics and Control Systems, no. 3(37), pp. 124–132, 2013.

[28] B. T. Polyak, "Convergence and convergence rate of iterative stochastic algorithms, I: General case," Autom. Remote Control, vol. 12, pp. 1858–1868, 1976.

[29] G. C. Goodwin, and K. S.Sin, Adaptive Filtering, Prediction and Control Engewood Cliffs, NJ.: Prentice-Hall, 1984.