SHADOWING PROPERTY IN ANALYSIS OF NEURAL NETWORKS DYNAMICS

Andrzej Bielecki, Jerzy Ombach

IMUJ PREPRINT 2003/04

Andrzej Bielecki

Institute of Computer Science, Jagiellonian University, Kraków, Poland
uibielec@cyf-kr.edu.pl

Jerzy Ombach

Institute of Mathematics, Jagiellonian University, Kraków, Poland,
ombach@im.uj.edu.pl

Abstract

We show that the shadowing and the inverse shadowing, the notions considered in the theory of dynamical systems, may be successfully applied in the analysis of a multilayer neural networks learning process. Our main result is, that generically any such a process is robust. Implementation implications are discussed as well.

KEYWORDS: shadowing, inverse shadowing, robustness, multilayer neural network learning process, gradient differential equation, Euler method, gradient descent method, dynamical system.

1 Introduction

The theory of dynamical systems is often used as a model of a multilayer neural networks learning process, see for example [2], [15], [24] and [31] and Section 3 in this paper. In general, there are two ways of such processes modelling. Perhaps most naturally is to consider a discrete time dynamical system, i.e. a system which orbits $\{x_n\}_{n\in\mathcal{Z}}$ satisfy the following difference equation:

$$x_{n+1} = f(x_n),$$

where f is a given map. Yet, many researchers prefer continuous time dynamical systems (flows) ϕ generated by differential equations of the form

$$\dot{x} = F(x),$$

as they are often easier to analyse. In the latter case we encounter the problem, what features of the flow are shared by its discretization which itself is a discrete time dynamical system, and which actually describes the learning process. More generally, we are interested in such dynamical properties of the discretization which have relevant interpretation in the context of the learning process.

In the sequel we will consider the problem of robustness of some discretization f of a given flow ϕ which may be vividly expressed in two parts. The first one is whether an orbit of f calculated by a computer really corresponds to some genuine (true) orbit. The second question is whether every true orbit of f can be recovered, at least with a given accuracy by a computer. The first problem is in fact a question about the shadowing property of the system, called also the pseudo-orbits tracing property (abbr. POTP) established by Bowen ([8]) while the second one corresponds to the property known as the inverse shadowing. The theory of shadowing and inverse shadowing has been intensively developed and emerged to become a significant part of the qualitative theory of dynamical systems with many interesting results ([16], [17], [18], [22], [23], [26], [27]).

Gradient methods are commonly used in engineering computation of neural networks. Therefore, in this paper we analyse shadowing and inverse shadowing concepts in connection with Runge-Kutta methods applied to a gradient differential equation which models a multilayer networks training process ([24]).

In Section 2 basic definitions and theorems on shadowing and inverse shadowing are presented whereas in Section 3 a gradient dynamical system is analysed in a context of a neural network training process. In Section 4 we establish a theorem which says that a generic learning processes is robust i.e. it is both shadowing and inverse shadowing.

2 Robustness

This section contains basic definitions and some results needed in the sequel concerning the shadowing property and the inverse shadowing property. We refer to the Pilyugin's book [26] and the survey paper [23] for more details on the subject and the theory of dynamical systems.

In this paper we consider a compact and smooth manifold $\mathcal{M}$ with a Riemannian metric d. Let $f : \mathcal{M} \to \mathcal{M}$ be a diffeomorphism, $f \in \text{Diff}(\mathcal{M})$. By $O_f(x)$ we denote the orbit of a point $x \in \mathcal{M}$, i.e. the sequence $\{x_n\}_{n\in\mathcal{Z}} \subset \mathcal{M}$ such that $x_0 = x$ and $x_{n+1} = f(x_n)$ for all $n \in \mathcal{Z}$. Since f is invertible, $O_f(x) = \{f^n(x)\}_{n\in\mathcal{Z}}$, where f^n denotes the n-th iteration of the diffeomorphism f. As usual, $\mathcal{Z}$ denotes the set of integers.

A sequence $\{y_n\}_{n\in\mathcal{Z}} \subset \mathcal{M}$ is called a δ-pseudo-orbit of f if

$$d(f(y_n), y_{n+1}) \leq \delta,$$

for all $n \in \mathcal{Z}$.

Definition 2.1 *The discrete time dynamical system generated by f (or just f) is shadowing, if for every $\varepsilon > 0$ there exists $\delta > 0$ such that any δ-pseudo-orbit $\{y_n\}_{n\in\mathcal{Z}}$ of the diffeomorphism f is ε-traced by the orbit of some point $x \in \mathcal{M}$, i.e.*

$$d(y_n, f^n(x)) \leq \varepsilon,$$

for all $n \in \mathcal{Z}$.

Let $\mathcal{M}^{\mathcal{Z}}$ denote the family of all sequences of elements of $\mathcal{M}$ indexed by $\mathcal{Z}$. Let us recall the concept of δ-method introduced by Kloeden and Ombach ([16]).

Definition 2.2 *A map $\mu_f : \mathcal{M} \longrightarrow \mathcal{M}^{\mathcal{Z}}$ is called a δ-method of the diffeomorphism f, if the following conditions hold:*

1. *$\mu_f(y)_0 = y$, for all $y \in \mathcal{M}$*
2. *$\mu_f(y)$ is a δ-pseudo-orbit of the map f.*

Example 1.

Let $g : \mathcal{M} \to \mathcal{M}$ be onto map and satisfy $D_\infty(f, g) \leq \delta$, where $D_\infty(f, g) \stackrel{def}{=} \sup_{x \in \mathcal{M}} d(f(x), g(x))$. We define the map

$$\mu_f(y) = O_g(y), \quad \text{for all } y \in \mathcal{M}.$$

Then μ_f is a δ-method of f.

Example 2.

Let $\chi_n : \mathcal{M} \to \mathcal{M}$, $n \in \mathcal{Z}$ be a family of maps such that $\chi_0 = id_{\mathcal{M}}$ and for all n, $D_\infty(f \circ \chi_n, \chi_{n+1}) \leq \delta$ and let:

$$\mu_f(y) = \{\chi_n(y)\}_{n \in \mathcal{Z}}, \quad \text{for all } y \in \mathcal{M}.$$

Then μ_f is a δ-method of f.

Example 3.

Let $\chi_n : \mathcal{M} \to \mathcal{M}$, $n \in \mathcal{Z}$ be a family of maps such that for all n, $D_\infty(f, \chi_n) \leq \delta$ and let:

$$\mu_f(y) = \{y_n\}_{n \in \mathcal{Z}} \text{ such that } y_0 = y,\ y_{n+1} = \chi_n(y_n) \ \text{ for all } y \in \mathcal{M}.$$

Then μ_f is a δ-method of f.

There are various approaches to introduce the concept of the inverse shadowing. Let us define it in the most general way. Denote by $\mathcal{T} = \mathcal{T}(f)$ a collection of δ-methods of f satisfying the condition: for any positive δ there is a δ-method $\mu_f \in \mathcal{T}$. Such $\mathcal{T}$ will be called a class. The set of all δ-methods is then a class and it will be denoted by $\mathcal{T}_0$.

The following four classes of δ-methods have been recently considered.

1. The class $\mathcal{T}_c(f)$ of all continuous δ-methods, where the continuity of a δ-method μ is defined with respect to the product topology in $\mathcal{M}^{\mathcal{Z}}$ ([16]).

2. The class $\mathcal{T}_h(f)$ of all δ-methods of such form as in Example 1, where the g's are homeomorphisms ([16]).
3. The class $\Theta_c(f)$ of all δ-methods of such form as in Example 2, where each χ_n is a continuous map ([27]).
4. The class $\Theta_s(f)$ of all δ-methods of such form as in Example 3, where each χ_n is a continuous map ([27]).

Let $\mathcal{T}$ be a class of δ-methods.

Definition 2.3 *The discrete time dynamical system generated by f (or just f) is $\mathcal{T}$ inverse shadowing if for any $\varepsilon > 0$ there is $\delta > 0$ such that for any orbit $\{x_n\}_{n\in\mathcal{Z}}$ and any δ-method $\mu_f \in \mathcal{T}$ there is $y \in \mathcal{M}$ such that*

$$d(x_n, \mu_f(y)_n) < \varepsilon,$$

for all $n \in \mathcal{Z}$.

Definition 2.4 *The discrete time dynamical system generated by f (or just f) is $\mathcal{T}$ robust (or bishadowing), if it is both shadowing and $\mathcal{T}$ inverse shadowing.*

Remarks

1. If $\mathcal{T} = \mathcal{T}_0$ then the above definition is the same as the concept of the inverse shadowing introduced and examined by Corless and Pilyugin in [9]. Actually, they did not use the notion of a δ-method there. In the same paper the authors showed that the definition was of a limited interest. In fact, they showed that any structurally stable diffeomorphism is not $\mathcal{T}_0$ inverse shadowing.
2. It is obvious that if $\mathcal{T}_1(f) \subset \mathcal{T}_2(f)$ and f is $\mathcal{T}_2(f)$ inverse shadowing then it is $\mathcal{T}_1(f)$ inverse shadowing as well.
3. $\mathcal{T}_h(f) \subset \Theta_c(f) \cap \Theta_s(f) \subset \Theta_c(f) \cup \Theta_s(f) \subset \mathcal{T}_0(f)$ and all the inclusions are proper. In particular, $\Theta_c(f)$ and $\Theta_s(f)$ do not include each other (see [27]).
4. We have $\Theta_c(f) = \mathcal{T}_c(f)$. Namely, for a given δ-method $\mu \in \mathcal{T}_c(f)$ one can define maps χ_n as $\chi_n(y) = \mu(y)_n$ and this means that $\mathcal{T}_c(f) \subset \Theta_c(f)$. The other inclusion is obvious.

Robustness is a topological conjugacy invariant in the following sense.

We say that diffeomorphisms $f,\ g : \mathcal{M} \longrightarrow \mathcal{M}$ are topologically conjugate if there exists a homeomorphism $\alpha : \mathcal{M} \longrightarrow \mathcal{M}$ such that

$$f \circ \alpha = \alpha \circ g \tag{1}$$

It appears that discrete time dynamical systems defined by topologically conjugate diffeomorphisms share all the properties which are expressed in topological terms. In particular we immediately have:

Theorem 2.5 *Let $f,\ g : \mathcal{M} \longrightarrow \mathcal{M}$ be topologically conjugate diffeomorphisms. For the class $\mathcal{T} = \Theta_c, \Theta_s$, $\mathcal{T}(f)$ robustness of f is equivalent to $\mathcal{T}(g)$ robustness of g.*

For our purposes we will use the well known concept of the Morse-Smale diffeomorphisms and the Morse-Smale flows. We recall the basic definition and for the details refer to the standard books on dynamical systems, see for example [25],[26].

Definition 2.6 *$f \in$ Diff$(\mathcal{M})$ is said to be a Morse-Smale diffeomorphism provided that its nonwandering set is a finite set of periodic orbits, each of which is hyperbolic and their stable and unstable manifolds are all transversal to each other.*

We have the following.

Theorem 2.7 *For the class $\mathcal{T} = \Theta_c(f), \Theta_s(f)$ any Morse-Smale diffeomorphism is $\mathcal{T}$ robust.*

Proof. In fact, the above theorem is an immediate consequence of the three fundamental results:

- any Morse-Smale diffeomorphism is structurally stable (see for example [25], p. 154),
- any structurally stable diffeomorphism is shadowing ([29]),
- any structurally stable diffeomorphism is both Θ_c and Θ_s inverse shadowing ([27]).

$\square$

We turn to the flows for a while. Given a C^1 vector field F on $\mathcal{M}$ we have a corresponding continuous time dynamical system (flow) generated by the equation $\dot{x}= F(x)$.

Definition 2.8 *F is said to be a Morse-Smale vector field provided that the nonwandering set of its flow is a finite union of periodic orbits and critical points of F, each of which is hyperbolic and their stable and unstable manifolds are all transversal to each other.*

These vector fields will reappear in Section 4.

3 Learning Process of a Neuron

In this section we summarize some basic concepts and results on the learning process of multilayer artificial neural networks. We refer to the book [15] and the survey papers [4], [7] for more information on the subject.

The mathematical theory which is the basis of the analysis of a training process is to some extent related to the concept of topological conjugacy (see previous section) of discretizations generated by a differential equation. More information on the subject and proves of theorems can be found in [3], [6], [10], [11], [12], [13], [14], [19].

There are several methods of learning of artificial neural networks. Most of them are iterative processes. One of the possible approaches to analysis of these processes is to consider differential equations such that the actual iterative procedure is a numerical method applied to this equation. In the paper [24] a gradient differential equation is proposed as a proper one for such analysis (see also discussion in [15], subsection 6.2). Indeed, let us notice that the descent gradient method which leads to the variation of synapses of the form

$$\vec{w}^{(p+1)} = \vec{w}^{(p)} - \eta \cdot grad\ E(\vec{w}^{(p)}), \tag{2}$$

where $\vec{w} = [w_1, ..., w_k]$ is a vector of weights of a neuron, is an iterative process generated by the Euler method for the differential equation

$$\dot{\vec{w}}= -grad\ E(\vec{w}). \tag{3}$$

If we consider a neuron, an output deviation function, called also a criterial function, plays a role of the potential E in the gradient equation (3). Most often the square criterial function is used, and for a single neuron it is given by the formula

$$E(\vec{w}) = \frac{1}{2} \sum_{i=1}^{N} \left[f(\beta^{(i)}) - z^{(i)} \right]^2 , \tag{4}$$

where f is an activation function of a neuron, $\beta^{(i)} := x_1^{(i)} \cdot w_1 + ... + x_k^{(i)} \cdot w_k$ and a vector $\vec{x}^{(i)} = [x_1^{(i)}, ..., x_k^{(i)}]$ is an input signal. A number $z^{(i)}$ is a desired response of the neuron if the vector $\vec{x}^{(i)}$ is given to the input and N is a number of input vectors used in the learning process. The finite sequence $\left((\vec{x}^{(1)}, z^{(1)}), ..., (\vec{x}^{(N)}, z^{(N)}) \right)$ is called the learning sequence.

The Euler method is a Runge-Kutta method of order $k = 1$. The Runge-Kutta methods of orders $k > 1$ are also sometimes considered as learning processes ([15]).

Moreover, although the gradient system (3) and its discretizations are defined formally on the whole space $\mathbf{R}^n$ with an appropriate n, we can confine ourselves to a finite dimensional compact manifold what simplifies the study.

Note first, that the range of numbers which can be represented in a computer is bounded. Furthermore, it should be stressed that in a biological neural cell, neurotransmitters are liberated in tiny amounts from vesicles - about 10^{-17} mol acetylocholin per impulse ([21], p.5, [30], pp.39-40). Thus, both in biological and artificial neural networks, absolute values of vectors $\vec{w}$ and $\vec{x}$ are bounded and, therefore, in modelling a neuron numerically we can consider only bounded vectors $\vec{w}$ and $\vec{x}$. It means that we are interested in the dynamics restricted to some set, possibly large but bounded. Let assume this set to be the ball $B(0, r)$ with some $r > 0$..

Now, we can assume the activation functions of each neuron to be bounded mappings of a class $\mathcal{C}^2(\mathbf{R}, \mathbf{R})$ with the first and second derivatives bounded as well. Most types of activative functions used in practice (for instance bipolar and unipolar ones and most radial functions) satisfy these assumptions. On the other hand, this makes it possible to modify the criterial function (4) in such a way that it is unchanged on the above ball $B(0, r)$ and the resulting system

$$\dot{x} = -grad\ E^*(x), x \in \mathbf{R}^n \tag{5}$$

is equivalent to a gradient system on the sphere S^n.

$$\dot{x} = -grad\ E^{**}(x), x \in S^n \tag{6}$$

A compactification-type construction is presented in [1], [2] and [5].

As it has been already mentioned, if one considers mathematical models of a training process, one of the basic questions is whether the qualitative properties of a continuous time system are preserved under an implementation. This question may be answered in terms of topological conjugacy between a discrete dynamical system obtained via the time discretization of the flow and a system generated by an applied numerical method.

It appears that on a finite-dimensional compact manifold $\mathcal{M}$ a gradient dynamical system is, under some natural assumptions, correctly reproduced by the Runge-Kutta method of order at least two for a sufficiently small time step. For the Euler method, which is the Runge-Kutta method of order one, the corresponding result has been proved only for a two-dimensional manifold. This means that the time-η-map of the induced dynamical system is topologically conjugate to the discrete dynamical system obtained via the Runge-Kutta method. This can be formally presented as follows.

Theorem 3.1 *Let $\mathcal{M}$ be a finite dimensional compact smooth manifold without a boundary and*

$$\phi : \mathcal{M} \times \mathbf{R} \longrightarrow \mathcal{M}$$

be a flow generated by a differential equation

$$\dot{x} = -grad\ E(x), \tag{7}$$

on the manifold $\mathcal{M}$, where $E \in \mathcal{C}^2(\mathcal{M}, \mathbf{R})$. Assume that ϕ has only a finite number of singularities, all of them hyperbollic, and there are not saddle-saddle connections. Denote by $\phi_\eta : \mathcal{M} \longrightarrow \mathcal{M}$ the time-η-map of the system ϕ, i.e. $\phi_\eta(x) := \phi(x, \eta)$, and by $\psi_{\eta,k}$ the diffeomorphism generated by the Runge-Kutta method of order k applied to equation (7). Assume also that if $k = 1$ then the manifold $\mathcal{M}$ is two-dimensional.

Then, for sufficiently small $\eta > 0$, there exists a homeomorphism $\alpha = \alpha_\eta : \mathcal{M} \longrightarrow \mathcal{M}$ conjugating discrete time dynamical systems *generated by ϕ_η and $\psi_{\eta,k}$, i.e. the following formula holds, compare with* (1)*:*

$$\psi_{\eta,k} \circ \alpha = \alpha \circ \phi_\eta. \tag{8}$$

The proof in the case if $k = 1$ and $\mathcal{M}$ – the two-dimensional sphere $\mathcal{S}^2$ is presented in [3]. The general version of the theorem for $k = 1$ is proved in [6]. The theorem for $k \geq 2$ is presented in [19]. Similar theorems for methods of order at least two has been investigated intensively ([10], [11], [12], [13], [14]).

It is worth to stress here, that under assumptions of the above Theorem all dynamical properties of the time-η-map ϕ are shared by $\psi_{\eta,k}$ while ϕ_η itself shares a lot of dynamical properties of the gradient flow ϕ. In particular, both ϕ_η and $\psi_{\eta,k}$ have finite many fixed points and each orbit converges to one of them.

Note, that if we want to use the above theorem to the study of the learning process described by the Euler method (called also the descent gradient method) which is used very widely, we are restricted to the case of a two-dimensional manifold. Therefore we can consider only a one-layer artificial neural network consisting of nonlinear neurons having two-componental inputs. Since in a one-layer artificial neural network neurons learn independently, we can consider the learning process only for a single neuron, which simplifies the situation. If we consider learning processes based on the Runge-Kutta methods of order $k > 1$, then we can consider any mulilayer network.

4 Robustness of the learning process

In this section we want to formulate and then justify the fact that a typical (generic) learning process is robust i.e. it is both shadowing and inverse shadowing with respect to a broad class of δ-methods.

In the theory of topological dynamical systems the word "typical" referrs to the property which is shared by systems from a large set, most often from so called a residual set. Here, we will use the word "typical" in its even stronger meaning. Namely, we will show that robustness is shared by learning processes resulting from vector fields belonging to some open and dense set in an appropriate space.

The argumentation in the previous Section makes it possible to identify each learning process with a Runge-Kutta method of a gradient differential equation of the form (6), where $E : \mathcal{M} \longrightarrow \mathbf{R}$ is a function defined on a compact manifold $\mathcal{M}$.

Denote by Γ the set of all C^1 vector fields on $\mathcal{M}$ equipped with the C^1 topology and let $\mathcal{G} \subset \Gamma$ be formed by all vector fields of the form $-grad\, E$, where $E : \mathcal{M} \longrightarrow \mathbf{R}$ is a C^2 function. With any vector field in $\mathcal{G}$ we associate its discretizations: time-η-map ϕ_η and Runge-Kutta methods $\psi_{\eta,k}$.

Theorem 4.1 *Let $\mathcal{T}$ be one of the classes $\mathcal{T}_h, \Theta_c, \Theta_s$. There exists an open and dense set of vector fields contained in $\mathcal{G}$ such that for a sufficiently small η*

1. *ϕ_η is $\mathcal{T}$ robust,*
2. *If $\dim \mathcal{M} = 2$, then $\psi_{\eta,1}$ is $\mathcal{T}$ robust.*
3. *If $\dim \mathcal{M} \geq 2$, then for each $k \geq 2$ $\psi_{\eta,k}$ is $\mathcal{T}$ robust.*

Proof. Denote by $MS_{\mathcal{G}}$ the set of all Morse-Smale vector fields contained in $\mathcal{G}$. The classical result is that the set $MS_{\mathcal{G}}$ is open and dense in $\mathcal{G}$, see for example [25], p. 153.

On the other hand, if $-grad\ E$ belongs to $MS_{\mathcal{G}}$, then the critical points of ϕ coincide with the fixed points of ϕ_η, and ϕ_η as ϕ does not admit other periodic orbits. Besides, stable and unstable manifolds of ϕ and ϕ_η at their (common) fixed points are the same. Hence, ϕ_η is a Morse-Smale diffeomorphism and by Theorem 2.7, and Remarks 2 and 3 in Section 2, is $\mathcal{T}$ robust, which proves 1.

Also, one can easily see that for $-grad\ E \in MS_{\mathcal{G}}$ all the assumptions of Theorem 3.1 are satisfied. Thus, ϕ_η and $\psi_{\eta,k}$ are topologically conjugate to each other if η is small enough, hence by Theorem 2.5 in Section 2 we have also proved 2 and 3. $\square$

At last we have to admit that the above result has some practical disadvantage. Namely, the classes $\mathcal{T}$ of δ-methods considered above, although quite large, do not contain real computer methods as the latter are only piecewise continuous. To be more specific, we would like to know that the learning process described above as ϕ_η or $\psi_{\eta,k}$ is inverse shadowing with respect to the class generated by real numerical methods like

$$\{\psi_{\eta,k,m},\ m \in \{1,2,...\}\} \tag{9}$$

where the subscript n is responsible for the round-off with set up, say 2^{-m}, accuracy. Such methods are piecewise constant, thus admit points of discontinuity and our framework does not work.

5 Concluding remarks

The dynamics of learning processes of some artificial and nonlinear neural networks can be understood using the theory of smooth dynamical systems and in many situations the gradient dynamical systems are the good tool for that. It appears that generically such processes are convergent to equilibrium states and are both shadowing and inverse shadowing. It means their robustness and may mean good enough accuracy while they are performed by a computer. The further studies on the inverse shadowing with respect to a class of piecewise continuous methods are still welcomed.

References

[1] Bielecki A., *Topological conjugacy of gradient cascades and learning process of a nonlinear neuron,* Opuscula Mathematica, vol.20, 2000, 19-25.

[2] Bielecki A., *Dynamical Properties of learning process of weakly nonlinear and nonlinear neurons,* Nonlinear Analysis: Real World Applications, vol.2, 2001, 249-258.

[3] Bielecki A., *Topological conjugacy of cascades generated by gradient flow on a two-dimensional sphere,* Annales Polonici Mathematici, vol.73, 2000, 37-57.

[4] Bielecki A., *Mathematical model of artificial neural networks,* in: Advances in Multi-Agent Systems, Schaefer R., Sędziwy S. - eds, Wydawnictwo Uniwersytetu Jagiellońskiego, Kraków, 2001.

[5] Bielecki A., *Agent's adaptivity - some aspects of theoretical foundations of a neural agent training process,* Lecture Notes in Artificial Intelligence, vol.2296, 47-53.

[6] Bielecki A., *Topological conjugacy of discrete-time map and Euler discrete dynamical systems generated by gradient flow on a two-dimensional compact manifold,* Nonlinear Analysis - Theory, Methods and Applications 51(2002), 1293 - 1317.

[7] Bielecki A., *Mathematical model of architecture and learning process of artificial neural networks,* Task Quarterly, vol.7, 2003, 93-114.

[8] Bowen R., *ω-limit sets for axiom a diffeomorphisms,* J. Diff. Equations, vol.18, 1975, 333–339.

[9] Corless R., Pilyugin S.Yu., *Approximate and real trajectories for generic dynamical systems,* J. Math. Anal. Appl., vol.189, 1995, 409–423.

[10] Garay B., *Discretization and some qualitative properties of ordinary differential equations about equilibria,* Acta Math. Univ. Comenianae 62 (1993), 245-275.

[11] Garay B., *Discretization and Morse-Smale dynamical systems on planar discs,* Acta Math. Univ. Comenianae 63 (1994), 25-38.

[12] Garay B., *On structural stability of ordinary differential equations with respect to discretization methods,* J. Numer. Math. 4 (1996), 449-479.

[13] Garay B., *On C^j-closeness between the solution flow and its numerical approximation,* J. Diff. Eq. App. 2 (1996), 67-86.

[14] Garay B., *The discretized flow on domains of attraction: a structural stability result,* IMA J. Numer. Anal. 18 (1998), 77-90.

[15] Hertz J., Krogh A., Palmer R.G., *Introduction to the theory of neural computation,* Addison-Wesley Publishing Company, Reedwood City, 1991.

[16] Kloeden P.E., Ombach J., *Hyperbolic homeomorphisms are bishadowing,* Annales Polonici Mathematici, vol.65, 1997, 171–177.

[17] Kloeden P.E., Ombach J., Pokrowskii A., *Continuous and inverse shadowing,* J. Func. Differential Equations, vol.6, 1999, 135-151.

[18] Lewowicz J., *Persitence in expansive systems,* Ergodic Theory and Dyn. Systems, 1983, 657–578.

[19] Li M.C., *Structural Stability of Morse-Smale Gradient-Like Flows under Discretization,* SIAM J. Math. Anal. 28, no.2 (1997), 381-388.

[20] Mañé R., *A Proof of C^1 Stability Conjecture,* Publ. Mathematiques IHES 66, 1988.

[21] Müller B., Reinhardt J., *Neural Networks,* Springer Verlag, New York, 1990.

[22] Ombach J., *The simplest shadowing,* Annales Polonici Mathematici, vol.58, 1993, 243-158.

[23] Ombach J., Mazur M., *Shadowing and likes as C^0 generic properties,* in W. Kryszewski, A. Nowakowski (Eds.) Proceedings of the 3rd Polish Symposium on Nonlinear Analysis, Lecture Notes in Nonlinear Analysis 3, Juliusz Schauder Center for Nonlinear Studies, Nicholas Copernicus University, 159 - 189, 2002.

[24] Owens A.J., Filkin D.L., *Efficient training of the back propagation network by solving a system of stiff ordinary differential equation,* International Joint Conference on Neural Networks, Washington, New York: IEEE, 1989, vol.II, 381-386.

[25] Palis J., de Melo W., *Geometric Theory of Dynamical Systems,* Springer Verlag, New York, 1982.

[26] Pilyugin S. Yu., *Shadowing in Dynamical Systems,* Lecture Notes in Mathematics 1706, Springer, 1999.

[27] Pilyugin S. Yu., *Inverse shadowing by continuous methods,* Discrete Contin. Dyn. Syst. 8(2000), 29-38.

[28] S. Yu. Pilyugin and O. B. Plamenevskaya, *Shadowing is generic*, Topology Appl. (1999), 253–266.

[29] Robinson C., *Stability theorems and hyperbolicity in dynamical systems,* Rocky Mountains J. of Math., vol.7, 1977, 425–437.

[30] Tadeusiewicz R., *Problemy biocybernetyki (Biocybernetics problems),* PWN, Warszawa, 1994 (in Polish).

[31] Wu W., Xu Y., *Deterministic convergence of an online gradient method for neural networks,* Journal of Computational and Applied Mathematics, vol.144, 2002, 335-347.