J Nonlinear Sci (2011) 21:579–593
DOI 10.1007/s00332-011-9094-1

Journal of
Nonlinear
Science

Dynamical Properties of a Perceptron Learning Process: Structural Stability Under Numerics and Shadowing

Andrzej Bielecki · Jerzy Ombach

Received: 26 March 2009 / Accepted: 4 February 2011 / Published online: 17 March 2011

Abstract In this paper two aspects of numerical dynamics are used for an artificial neural network (ANN) analysis. It is shown that topological conjugacy of gradient dynamical systems and both the shadowing and inverse shadowing properties have nontrivial implications in the analysis of a perceptron learning process. The main result is that, generically, any such process is stable under numerics and robust. Implementation aspects are discussed as well. The analysis is based on the theorem concerning global topological conjugacy of cascades generated by a gradient flow on a compact manifold without a boundary.

Keywords Dynamical system · Topological conjugacy · Shadowing · Inverse shadowing · Robustness · Perceptron learning process · Gradient differential equation · Runge–Kutta methods

Mathematics Subject Classification (2000) Primary 37C15 · 37C50 · 65L06 · 68T05 · Secondary 34D30 · 37C10 · 37C20 · 37M99 · 37N30 · 65L20 · 68Q32

Communicated by Peter Kloeden.

A. Bielecki (✉)
Institute of Computer Science, Jagiellonian University, Łojasiewicza 6, 30-348 Kraków, Poland
e-mail: bielecki@ii.uj.edu.pl

J. Ombach
Institute of Mathematics, Jagiellonian University, Łojasiewicza 6, 30-348 Kraków, Poland
e-mail: Jerzy.Ombach@im.uj.edu.pl

1 Introduction

The analysis of the learning process properties of a multilayer artificial neural network (ANN), also called a perceptron after the name of the first multi-layer ANN implemented by Rosenblatt (1961), is a classical example of an application of dynamical systems theory to the analysis of neural networks properties—see for instance Bielecki (2001), Bielecki and Ombach (2004), Hertz et al. (1991), Wu and Xu (2002). The most natural approach is to consider a cascade, generated by a numerical process, which is used for an ANN learning process realization. Gradient one-step methods are commonly used in engineering computation of neural networks as learning algorithms of perceptrons. According to the mentioned methodology we consider both the problem of topological conjugacy and shadowing. In Sect. 2 basic definitions and theorems on topological conjugacy, shadowing, and inverse shadowing are recalled, whereas in Sect. 3 a formal approach to a perceptron learning process analysis is presented.

This paper is a continuation of studies presented in Bielecki (2001) and Bielecki and Ombach (2004) where, based on results obtained in Bielecki (2002), the stability of a learning process of a neuron having two-componental input was proved (Bielecki 2001), and a bishadowing property (robustness) of a perceptron learning process in the case when a Runge–Kutta method of order at least two was established (Bielecki and Ombach 2004). For the Euler method, the most common one used for perceptron learning and also called the gradient descent method, which is the Runge–Kutta method of order one, the robustness and stability analysis of a learning process of only a two-componental input single neuron were considered (Bielecki 2001; Bielecki and Ombach 2004). It was related to the fact that the theorem about topological conjugacy, on which the analysis was based, had been proved only for a two-dimensional manifold. Considering the fact that Runge–Kutta methods of order greater than two are not used at all as ANN learning algorithms, and those of order two are used rarely whereas of order one widely, the analysis was very incomplete. Furthermore, the applied method based on compactification via stereographic projection has one additional disadvantage: the obtained results cannot be applied directly to cascades on $\mathbb{R}^n$. Therefore, in this paper a different method of compactification is introduced; see Sect. 3. The analysis described here is based on the result obtained by Li (1999), who proved topological conjugacy for any finite-dimensional compact manifold, but the result has a slightly different form than commonly used ones in this context. See Theorem 2.5, (3) and Remark 2. However, this result allows us to fill the mentioned gap concerning applications. The main result of this paper, Theorem 3.1, states that the perceptron learning process is, generically, both stable under numerics and robust according to every Runge–Kutta method, including the gradient descent method, which is widely used as a perceptron training algorithm. In order to apply theorems concerning cascade properties on a manifold, a special manifold, resembling (in a three-dimensional case) a round mattress laying on a plane, is constructed—see Step 1 of Theorem 3.1 and Fig. 1.

2 Mathematical Foundations

Throughout the paper cascades on a compact, smooth, Riemannian manifold $\mathcal{M}$ without boundaries are considered. Let us denote by ϱ the Riemannian metric on $\mathcal{M}$. In Sect. 2.1 definitions and theorems concerning topological conjugacy of cascades generated by a flow are presented. Foundations of shadowing and inverse shadowing theory are recalled in Sect. 2.2.

2.1 Topological Conjugacy

As has been mentioned, if one considers mathematical models of an ANN training process, the basic question is whether the qualitative properties of a continuous system are preserved under an implementation. Topological conjugacy is a standard tool for investigating the equivalence of dynamical systems according to their dynamics. For cascades defined by diffeomorphisms, topological conjugacy is defined in the following way.

Definition 2.1 We say that diffeomorphisms $f, g : \mathcal{M} \to \mathcal{M}$ are topologically conjugate if there exists a homeomorphism $\alpha : \mathcal{M} \to \mathcal{M}$ such that

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

In the sequel Morse–Smale dynamical systems are considered. Let us recall the basic definitions (for the details, refer to the standard books on dynamical systems, for example, Palis and de Melo 1982; Pilyugin 1999).

Definition 2.2 A mapping $f \in \mathrm{Diff}(\mathcal{M})$, or a cascade generated by this mapping, is said to be a Morse–Smale mapping provided that its nonwandering set is a finite set of periodic orbits and fixed points, each of which is hyperbolic and whose stable and unstable manifolds are all transversal to each other.

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 2.3 A flow is said to be a Morse–Smale flow provided that its nonwandering set is a finite union of periodic orbits and equilibrium points, each of which is hyperbolic and whose stable and unstable manifolds are all transversal to each other; i.e., the strong transversality condition is satisfied. Furthermore, there are no saddle–saddle connections. A vector field F is called a Morse–Smale vector field if it generates a Morse–Smale flow.

Definition 2.4 A dynamical system, both a cascade and a flow, is said to be Morse–Smale gradient-like provided that it is a Morse–Smale system having no periodic orbits.

Remark 1 Since a gradient dynamical system has no periodic orbits, each gradient system that is a Morse–Smale one is a Morse–Smale gradient-like system.

Recall the theorem proposed by Li (1999, Theorem 3).

Theorem 2.5 *Let $\mathcal{M}$ be a finite-dimensional compact smooth manifold without a boundary and let*

$$\phi : \mathcal{M} \times \mathbb{R} \to \mathcal{M}$$

be a Morse–Smale gradient-like flow, and denote it by $(\mathcal{M}, \phi)$, generated by a differential equation on the manifold $\mathcal{M}$

$$\dot{x} = F(x), \tag{2}$$

where F is a $\mathcal{C}^2$ vector field on $\mathcal{M}$. Denote by $\phi_h : \mathcal{M} \to \mathcal{M}$ the time-h-map of the system ϕ, i.e., $\phi_h(x) := \phi(x, h)$, and by ψ_h the diffeomorphism generated by the Euler method of the step size h applied to (2).

*Let $T > 0$ be given. Then, for sufficiently large m, there is a homeomorphism $\alpha_m : \mathcal{M} \to \mathcal{M}$ conjugating discrete-time dynamical systems generated by ϕ_T and the m*th *iteration $\psi^m_{\frac{T}{m}}$ of the operator $\psi_{\frac{T}{m}}$; i.e., the following formula holds—compare with (1):*

$$\psi^m_{\frac{T}{m}} \circ \alpha_m = \alpha_m \circ \phi_T. \tag{3}$$

Furthermore, $\lim_{m\to\infty} \varrho(\alpha_m(x), x) = 0$.

Remark 2

1. The above theorem for numerical methods of order at least two (i.e., $k \geq 2$) was proved by Li (1997) and also by Garay (1994) in a classical form, i.e., $m = 1$, $T = h$, where $h > 0$ is sufficiently small and α depends on h. In the method used there the conjugating homeomorphism was obtained by solving a certain functional equation. Note that the case of a manifold with a boundary is also considered in this paper. The proof for the Euler method on a two-dimensional compact manifold without a boundary for a gradient system, based on the estimation of accuracy of the Euler method on a Riemannian manifold (see Bielecki 2002), was presented in Bielecki (2002), also in a classical form. Local conjugacies were constructed using the basic domain method, and then they were glued. We stress that using the basic domain method allowed one to prove that stability under numerics also exists in a very particular case of saddle–saddle connection presence (Bielecki 2002, Lemma 5.2.1).
2. Let us notice that a classical form implies (3). Indeed, let us assume that there exists $h_0 > 0$ such that for each $0 < h < h_0$ the conjugating formula is satisfied:

   $$\psi_h \circ \alpha_h = \alpha_h \circ \phi_h.$$

   This implies

   $$\psi^m_h \circ \alpha_h = \alpha_h \circ \phi^m_h$$

   for any natural m. But $\phi^m_h = \phi(\cdot, mh)$, thus we obtain

   $$\psi^m_h \circ \alpha_h = \alpha_h \circ \phi(\cdot, mh).$$

Let $T > 0$ be given. Set $T = mh$, $0 < h < h_0$. Then α becomes a function of m and we have

$$\psi^m_{\frac{T}{m}} \circ \alpha_m = \alpha_m \circ \phi(\cdot, T).$$

Based on Li (1997, 1999), and point 2 of Remark 2, we can sum up the results by Li in the following form.

Theorem 2.6 *Let all assumptions concerning $\mathcal{M}$ and ϕ specified in Theorem 2.5 be satisfied. Denote by $\psi_{h,p}$ the diffeomorphism generated by a Runge–Kutta method of the step size h and order p applied to* (2). *Then, for sufficiently large m and each* $p \in \{1, 2, \ldots\}$, *there is a homeomorphism $\alpha_m : \mathcal{M} \to \mathcal{M}$ such that*

$$\psi^m_{\frac{T}{m},p} \circ \alpha_m = \alpha_m \circ \phi_T.$$

Furthermore, $\lim_{m\to\infty} \varrho(\alpha_m(x), x) = 0$.

A flow is stable according to a numerical method if cascades generated by this method and time discretization have the same dynamical properties. Formally, it is defined in the following way.

Definition 2.7 Let a numerical method applied to the flow generated by (2) be given by the operator $\Psi : \mathcal{M} \to \mathcal{M}$. A flow ϕ is stable under numerics with respect to the operator Ψ if cascades generated by the time discretization of the flow ϕ and the operator Ψ are topologically conjugate.

In the theory of topological dynamical systems the word "typical" refers to the property which is shared by systems from a large set, most often from what is called a residual set. Here, we will use the word "typical" in an even stronger meaning. It turns out that systems satisfying the assumptions of Theorem 2.5 are typical in the space of gradient dynamical systems (see Sect. 3) in the strongest meaning; i.e., assumptions are generic according to the following definition.

Definition 2.8 A given property is said to be generic in a topological space X if there exists an open and dense set in X having this property.

Theorem 2.6 implies 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 for a sufficiently small time step. This fact with a few implications can be used as the formal foundations of a perceptron learning process analysis. Section 3 presents an analysis of a perceptron learning process stability under numerics that is based, among other things, on Theorem 2.6.

2.2 Shadowing

This section contains basic definitions and some results needed in the sequel concerning both the shadowing and the inverse shadowing properties. We refer to Pilyugin's book (Pilyugin 1999) for more details on the subject and on the theory of dynamical systems.

Let $f : \mathcal{M} \to \mathcal{M}$ be a diffeomorphism, $f \in \mathrm{Diff}(\mathcal{M})$. By $O_f(x)$ we denote the orbit of a point $x \in \mathcal{M}$, i.e., the sequence $\{x_k\}_{k\in\mathbb{Z}} \subset \mathcal{M}$ such that $x_0 = x$ and $x_{k+1} = f(x_k)$ for all $k \in \mathbb{Z}$. Since f is invertible, $O_f(x) = \{f^k(x)\}_{k\in\mathbb{Z}}$.

Definition 2.9 A sequence $\{y_k\}_{k\in\mathbb{Z}} \subset \mathcal{M}$ is called a δ-pseudo-orbit of f if

$$\varrho\big(f(y_k), y_{k+1}\big) \le \delta,$$

for all $k \in \mathbb{Z}$.

Definition 2.10 The discrete-time dynamical system generated by f is shadowing, if for every $\varepsilon > 0$ there exists $\delta > 0$ such that any δ-pseudo-orbit $\{y_k\}_{k\in\mathbb{Z}}$ of the diffeomorphism f is ε-traced by the orbit of some point $x \in \mathcal{M}$, i.e.,

$$\varrho\big(y_k, f^k(x)\big) \le \varepsilon,$$

for all $k \in \mathbb{Z}$.

Let $\mathcal{M}^{\mathbb{Z}}$ denote the family of all sequences of elements of $\mathcal{M}$ indexed by $\mathbb{Z}$. Let us recall the concept of δ-method introduced by Kloeden and Ombach (1997).

Definition 2.11 A map $\mu_f : \mathcal{M} \to \mathcal{M}^{\mathbb{Z}}$ 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.

There are various approaches to introduce the concept of 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 following 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. Examples of some other classes and their properties can be found in Bielecki and Ombach (2004).

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

Definition 2.12 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 $\{x_k\}_{k\in\mathbb{Z}}$ and any δ-method $\mu_f \in \mathcal{T}$ there is $y \in \mathcal{M}$ such that

$$\varrho\big(x_k, \mu_f(y)_k\big) < \varepsilon,$$

for all $k \in \mathbb{Z}$.

Definition 2.13 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.

Remark 3 It is clear that the above defined robustness with respect to $\mathbb{Z}$ implies robustness with respect to $\mathbb{N}$.

In this paper we will use a class of methods that seems to be the largest one for which some results on inverse shadowing have been established until now. It is the union of two classes: $\Theta = \Theta_c \cup \Theta_s$. The class Θ_c consists of methods of the form

$$\mu_f(y) = \{\chi_k(y)\}_{k\in\mathbb{Z}}, \quad \text{for all } y \in \mathcal{M},$$

where $\chi_k : \mathcal{M} \to \mathcal{M}$, $k \in \mathbb{Z}$, is a family of continuous maps such that $\chi_0 = id_{\mathcal{M}}$ and, for all k, $D_\infty(f \circ \chi_k, \chi_{k+1}) \leq \delta$. The class Θ_s consists of methods of the form

$$\mu_f(y) = \{y_k\}_{k\in\mathbb{Z}} \quad \text{such that } y_0 = y,\ y_{k+1} = \chi_k(y_k) \ \text{ for all } y \in \mathcal{M},$$

where $\chi_k : \mathcal{M} \to \mathcal{M}$, $k \in \mathbb{Z}$ is a family of continuous maps such that $\chi_0 = id_{\mathcal{M}}$ and, for all k, $D_\infty(f \circ \chi_k, \chi_{k+1}) \leq \delta$. Here $D_\infty(g, h) := \sup_{x\in\mathcal{M}} d(g(x), h(x))$.

Robustness is a topological conjugacy invariant. In particular, we immediately have the following.

Theorem 2.14 *Let $f, g : \mathcal{M} \to \mathcal{M}$ be topologically conjugate diffeomorphisms. For the class $\mathcal{T} = \Theta$, $\mathcal{T}(f)$ robustness of f is equivalent to $\mathcal{T}(g)$ robustness of g.*

In order to prove Theorem 3.1, the following lemma, proved in Bielecki and Ombach (2004), will be used.

Lemma 2.15 *For the class $\mathcal{T} = \Theta$ any Morse–Smale diffeomorphism is $\mathcal{T}$ robust.*

3 Learning Process of a Perceptron

In this section we summarize some basic concepts and results on the learning process of multilayer artificial neural networks (ANNs). These kinds of ANNs are organized in such a way that the set of all neurons of which the perceptron is built can be decomposed into a finite family of disjoint finite sets $A_1, \ldots, A_U$ (layers), such that the output signal of each neuron belonging to the layer A_u is given to inputs of all neurons of the layer A_{u+1}, where $u \in \{1, \ldots, U - 1\}$. A neuron is a unit transforming an input signal $\vec{x}$ into the output signal $y = f(s)$, where $f : \mathbb{R} \to \mathbb{R}$ is an activation function of a neuron, and $s^{(i)} := x_1 \cdot w_1^{(i)} + \cdots + x_l \cdot w_l^{(i)}$ and $w_1^{(i)}, \ldots, w_l^{(i)}$ are weights (synapses) of the ith neuron. We refer to the book (Hertz et al. 1991) for more information on the subject of neuron models, perceptrons, and network learning processes. The mathematical theory which can be used as the basis of the analysis of perceptron gradient training methods is to some extent related to the concept of topological conjugacy and shadowing (see previous section) of discretizations generated by a differential equation.

There are several methods of ANN learning, and most of them are iterative processes. One of the possible approaches to analyze these processes is to consider differential equations such that the actual iterative procedure is a numerical method applied to this equation.

Considering the differential model of a learning process, let us notice that the gradient descent method leads to the iterative variation of synapses,

$$\vec{w}(k+1) = \vec{w}(k) - h \cdot \operatorname{grad} E\bigl(\vec{w}(k)\bigr), \tag{4}$$

where $\vec{w} = [w_1, \dots, w_n]$ is a vector of all weights of a perceptron, whereas k numerates a step of the learning process. The formula (4) describes the iterative process generated by the Euler method for the differential equation

$$\dot{\vec{w}} = -\operatorname{grad} E(\vec{w}). \tag{5}$$

An output deviation function E, also called a criterial function, plays the role of the potential E in the gradient equation (5). Equation (5) generates the gradient flow $(\mathbb{R}^n, \phi)$.

In order to explain in detail a perceptron learning process and, consequently, the meaning of the function E, assume that a finite sequence $((\vec{x}^{(1)}, \vec{z}^{(1)}), \dots, (\vec{x}^{(J)}, \vec{z}^{(J)}))$, called the learning sequence, is given, where $\vec{z}^{(j)}$ is a desired response of the perceptron if the vector $\vec{x}^{(j)}$ is put to its input and J is the number of input vectors used in the learning process. Since the real function E is a criterion of how correctly all weights of the perceptron are set, it should have nonnegative values and exactly one global minimum with a value equal to zero at the point $\vec{w}_0$ such that $\vec{y}^{(j)}(\vec{w}_0) = \vec{z}^{(j)}$ for each $j \in \{1, \dots, J\}$. Furthermore, the greater the differences between responses $\vec{y}^{(j)}$ of the perceptron and the proper responses $\vec{z}^{(j)}$, the greater the value of the function E. Assuming that the perceptron has n weights, the function $E : \mathbb{R}^n \to \mathbb{R}$ and, therefore, (5) generates a flow on the n-dimensional Euclidean space $\mathbb{R}^n$.

Most often the square criterial function is used, which is defined by the formula

$$E(\vec{w}) = \frac{1}{2} \sum_{j=1}^{J} \bigl[\vec{y}^{(j)}(\vec{w}) - \vec{z}^{(j)}\bigr]^2, \tag{6}$$

where $\vec{y}^{(j)}(\vec{w})$ is the output signal of the perceptron if the vector $\vec{x}^{(j)}$ is put to its input. Assuming that the activation function of each neuron is a mapping of the class $\mathcal{C}^2(\mathbb{R}, \mathbb{R})$—most types of activation functions used in practice, e.g., bipolar and unipolar sigmoid functions and most radial functions, satisfy this assumption—the criterial function E is also of the class $\mathcal{C}^2(\mathbb{R}^n, \mathbb{R})$. The formula (4) describes a process of finding a local minimum of the function E using the Euler method, which is a Runge–Kutta method of order $k = 1$. The Runge–Kutta methods of orders $k = 2$ are also sometimes considered as learning processes (Hertz et al. 1991). Moreover, although the gradient system (5) and its discretizations are defined formally on the space $\mathbb{R}^n$ with an appropriate n, we can study the learning process on the n-dimensional compact manifold, say $\mathcal{M}_S^n$, which is homeomorphic to the sphere $\mathcal{S}^n$, applying the compactification procedure. In this paper the procedure is an alternative to the one presented in Bielecki (2001) and Bielecki and Ombach (2004) because of its disadvantage mentioned in the introduction. The construction of the manifold $\mathcal{M}_S^n$ is described in detail in Step 1 of the proof of Theorem 3.1.

Let us consider the problem of the potential regularization. Denote by $B^n(\mathbf{0}, r)$ a closed, n-dimensional ball in $\mathbb{R}^n$, where $\mathbf{0}$ denotes zero in $\mathbb{R}^n$. In applications both the discrete and continuous models of a perceptron learning process (4) and (5) are considered in $\mathbb{R}^n$. In order to apply Theorem 2.6, the dynamical systems describing the learning process must be transformed onto a compact, smooth manifold without a boundary. In order to transform the learning process model (5) from $\mathbb{R}^n$ onto $\mathcal{M}_S^n$ let us modify the criterial function in such a way that on a certain ball, say $B^n(\mathbf{0}, r_1) \subset \mathbb{R}^n$, the potential is not modified—the radius r_1 can be as large as we need—and a ball $B^n(\mathbf{0}, 2r_1)$ will be an invariant set. Let, furthermore, $E(\vec{w}) = E(r)$, $r = \|\vec{w}\|^2$ (the square dependence is established because of the clarity of calculations; see Step 2 of the proof of Theorem 3.1) for r large but less than $2r_1$. In this way a flow $(B^n(\mathbf{0}, 2r_1), \tilde{\phi})$ is obtained. This procedure of the potential E regularization is described in detail below as the second step of Theorem 3.1.

Remark 4 Let us notice that this method of criterial function modification is well based on the properties on the modeled realities. Note first that the range of numbers which can be represented in a computer is bounded. Also, in a biological neural cell, neurotransmitters are liberated in tiny amounts from vesicles, about 10^{-17} mol per impulse (Hess 2009, Sect. 2.5, and Tadeusiewicz 1994, pp. 39–40). Thus, in both biological and artificial neural networks, the norms of vectors $\vec{w}$ and $\vec{x}$ are bounded; therefore, in modeling 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 us assume this set to be a ball $B^n(\mathbf{0}, r_1)$ with the radius r_1 sufficiently large.

Recapitulating, the criterial function (6) is unchanged on the above ball $B^n(\mathbf{0}, r_1)$, and the resulting system $(B^n(\mathbf{0}, 2r_1), \tilde{\phi})$ generated by the equation

$$\dot{\vec{w}} = -\operatorname{grad} \tilde{E}(\vec{w}), \qquad \vec{w} \in B^n(\mathbf{0}, R) \subset \mathbb{R}^n \tag{7}$$

is good for modeling of the learning process.

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 $-\operatorname{grad} E$, where $E : \mathcal{M} \to \mathbb{R}$ is a C^2 function. With any vector field in $\mathcal{G}$ we associate its discretizations: ϕ_T and Runge–Kutta methods $\psi_{\frac{T}{m},p}$. Let $\Psi = \psi^m_{\frac{T}{m},p}$ (see Definition 2.7) and $\mathcal{T} = \Theta$ (see Sect. 2.2). The dynamic properties of a learning process of a perceptron with n weights can be specified in the following way.

Theorem 3.1 *Fix a real number $T > 0$. Let a learning process of a perceptron having n weights be modeled by a flow $\tilde{\phi}$ on $B^n(\mathbf{0}, 2r_1) \subset \mathbb{R}^n$; see formula (7). Then there exists a compact, smooth, n-dimensional manifold $\mathcal{M}_S^n$ without a boundary and a flow $(\mathcal{M}_S^n, \hat{\phi})$ such that $B^n(\mathbf{0}, 2r_1) \subset \mathcal{M}_S^n$, $(\mathcal{M}_S^n | B^n(\mathbf{0}, 2r_1), \hat{\phi}) = (B^n(\mathbf{0}, 2r_1), \tilde{\phi})$, and the flow $(\mathcal{M}_S^n, \hat{\phi})$ is, generically, stable under numerics with respect to the operator Ψ, which means that cascades $(\mathcal{M}_S^n, \hat{\phi}_T)$ and $(\mathcal{M}_S^n, \Psi)$ are topologically conjugate. Furthermore, both the mentioned cascades $\hat{\phi}_T$ and Ψ are $\mathcal{T}$ generically robust as well.*

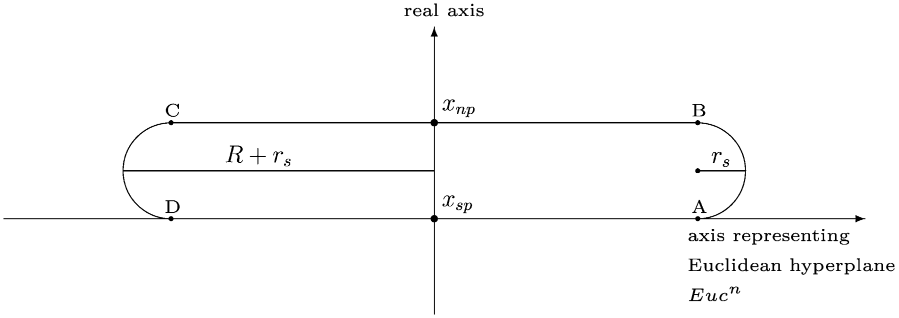

Fig. 1 Construction of the manifold $\mathcal{M}_S^n$

Proof

Step 1. Construction of the Manifold $\mathcal{M}_S^n$ Fix a closed ball $B^n(\mathbf{0}, r_1) \subset \mathbb{R}^n$, where the radius r_1 is as large as we need (see Remark 4) and set $R = 3r_1$. Let us construct a manifold $\mathcal{M}_S^n \in \mathbb{R}^n \times \mathbb{R}$ in such a way that it has a radial symmetry with respect to rotations around the real axis which is orthogonal to the Euclidean hyperplane, denoted by Euc^n, in which the mentioned ball $B^n(\mathbf{0}, R)$ is contained. See Fig. 1. Because of the radial symmetry, the construction can be described for each two-dimensional section. Thus, let us glue the line segment $[-R, R]$ with two hemicircles of a circle of radius r_s in the points $(-R, 0)$ and $(R, 0)$, respectively. Then glue the obtained curve with the line segment (see Fig. 1). The obtained manifold is compact, as it is homeomorphic to S^n. It is also of class $\mathcal{C}^1$. The lack of $\mathcal{C}^\infty$ smoothness in the points A, B, C, D on a two-dimensional section can be counterbalanced by a mollifier function, denoted by $f_{[a,b]}(x) \in \mathcal{C}^\infty(\mathbb{R})$. Let $f_{[a,b]}$ be of the form: $f_{[a,b]}(x) = 0$ for $x \in (-\infty, a]$, $f_{[a,b]}(x) = 1$ for $x \in [b, \infty)$ and $f_{[a,b]}$ is increasing on $[a, b]$. This type of function is called a cutoff function, and its construction is described, for instance, in Lee (2003, Lemma 2.21, p. 50). From the symmetry of the two-dimensional section, it is sufficient to describe the smoothing procedure only at the point A. We can treat the quarter of the section as the function $f_{\mathrm{sec}} : [0, R + r_s] \to [0, r_s]$ of the form

$$f_{\mathrm{sec}}(x) := \begin{cases} 0 & \text{for } r \in [0, R), \\ r_s - \sqrt{r_s^2 - (x-R)^2} & \text{for } [R, R + r_s]. \end{cases}$$

Then $A = (R, 0)$. Cut the domain of $f_{[R,R+\frac{r_s}{2}]}$ to the interval $[0, R + r_s]$ and set $f_{\mathrm{smooth}}(x) := f_{\mathrm{sec}}(x) \cdot f_{[R,R+\frac{r_s}{2}]}(x)$. It is easy to check that $f_{\mathrm{smooth}} \in \mathcal{C}^\infty(0, R + r_s)$. The manifold $\mathcal{M}_S^n$ is obtained by rotation of the two-dimensional section smoothed at the points A, B, C, and D around the real axis (see Fig. 1).

Denote by *Base* the part of $\mathcal{M}_S^n$ belonging to Euc^n, i.e., $Base := B^n(\mathbf{0}, R)$ and let $Cap := \mathcal{M}_S^n \setminus Base$. Notice that $B^n(\mathbf{0}, 2r_1) \subset Euc^n$, on which the dynamical system $\tilde{\phi}$ is founded, is a subset of *Base*.

Step 2. Compactification The training process will be considered on a closed ball $B^n(\mathbf{0}, r_1)$, and the potential outside $B^n(\mathbf{0}, r_1)$ will be modified and completed in order to apply theorems concerning properties of cascades considered on a compact manifold without boundaries. Thus, let us modify the potential E using a function g defined as follows:

$$g(\vec{w}) := \begin{cases} 1 & \text{for } r \in [0, r_1), \\ e^{(r-r_1)^a} & \text{for } r \geq r_1, \end{cases}$$

where $r := \|\vec{w}\|^2$ (a square dependence is chosen for clarity because then $\frac{\partial r}{\partial w_i} = 2w_i$ provided that $\|\cdot\|$ is the Euclidean norm) and a natural number a is selected depending on the potential E and radius r_1 in the way specified below. The function g is of a class $\mathcal{C}^2(\mathbb{R}^n, \mathbb{R})$ for $a > 2$. Define a potential $\tilde{E} : \mathbb{R}^n \to \mathbb{R}$ as $\tilde{E}(\vec{w}) := E(\vec{w}) \cdot g(\vec{w})$. For a sufficiently large a, solutions of the equation $\dot{\vec{w}} = -\operatorname{grad} \tilde{E}(\vec{w})$, generating a flow $\tilde{\phi}$, cut the $(n-1)$-dimensional sphere $\mathcal{S}^{n-1}(\mathbf{0}, 2r_1) \subset \mathit{Base}$ transversally, entering the interior of the ball $B^n(\mathbf{0}, 2r_1)$; that is, the scalar product $-\operatorname{grad} \tilde{E}(\vec{w}) \circ \vec{w}$ has negative values for $r = 2r_1$. This means that the ball $B^n(\mathbf{0}, 2r_1)$ is an invariant set of the flow $\tilde{\phi}$. Indeed, calculating the ith component of the scalar product $-\operatorname{grad} \tilde{E}(\vec{w}) \circ \vec{w}$ we obtain

$$\begin{aligned} -w_i \cdot \frac{\partial \tilde{E}(\vec{w})}{\partial w_i} &= -w_i \cdot \frac{\partial}{\partial w_i}\big(E(\vec{w}) \cdot g(\vec{w})\big) \\ &= -w_i \left(E(\vec{w}) \cdot \frac{\partial g(\vec{w})}{\partial w_i} + g(\vec{w}) \cdot \frac{\partial E(\vec{w})}{\partial w_i} \right) = \cdots . \end{aligned}$$

Because

$$\frac{\partial g(\vec{w})}{\partial w_i} := \begin{cases} 2 \cdot w_i \cdot a \cdot (r - r_1)^{a-1} \cdot e^{(r-r_1)^a} & \text{for } r > r_1, \\ 0 & \text{for } r \in [0, r_1], \end{cases}$$

then, continuing the calculation, we obtain for $r = 2r_1$

$$\begin{aligned} \cdots &= \left(-2w_i^2 a (r - r_1)^{a-1} E(\vec{w}) - w_i \frac{\partial E(\vec{w})}{\partial w_i} \right) e^{(r-r_1)^a} \\ &= \left(-2w_i^2 a r_1^{a-1} E(\vec{w}) - w_i \frac{\partial E(\vec{w})}{\partial w_i} \right) e^{r_1^a}. \end{aligned}$$

Thus, as on $\mathcal{S}^{n-1}(\mathbf{0}, 2r_1)$ we have $\sum_i w_i^2 := \|\vec{w}\|^2 = 2r_1$, so

$$-\operatorname{grad} \tilde{E}(\vec{w}) \circ \vec{w} = e^{r_1^a} \left(-4ar_1^a E(\vec{w}) - \sum_i w_i \frac{\partial E(\vec{w})}{\partial w_i} \right).$$

Since the problem is considered on the compact set $\mathcal{S}^{n-1}(\mathbf{0}, 2r_1)$, all variables, functions, and derivatives are bounded. In particular, the term $-\sum_i w_i \frac{\partial E(\vec{w})}{\partial w_i}$ can be positive, but is upper bounded. The potential E is nonnegative, and the flow (5) has only a finite number of singularities, which implies that E has only a finite number of zeros. Therefore, r_1 can be chosen so that $E(\vec{w}) > 0$ for each $\vec{w}$ such that $\|\vec{w}\|^2 = 2r_1 > 0$. Because the term $\sum_i w_i \frac{\partial E(\vec{w})}{\partial w_i}$ does not depend on a and r_1 is large, the number a can be chosen sufficiently large that $4ar_1^a E(\vec{w}) > |\sum_i w_i \frac{\partial E(\vec{w})}{\partial w_i}|$.

Cut the domain of $\tilde{E}$ to $B^n(\mathbf{0}, 2r_1) \subset Base$ and complete the potential on $\mathcal{M}_S^n$ so that in the north pole x_{np} (see Fig. 1) there is a hyperbolic fixed point which is a repeller on $\mathcal{M}_S^n \setminus B^n(\mathbf{0}, 2r_1)$, which means that all points in $\mathcal{M}_S^n \setminus B^n(\mathbf{0}, 2r_1)$ are attracted to the north pole in negative time. Then glue the potential C^2 regularly on the border of $B^n(\mathbf{0}, 2r_1)$. This can be done in the following way. Define the potential on $Cap \cup \partial Base$ as $V(\vec{w}) := c \cdot \varrho(x_{sp}, \vec{w})$, where $c > 0$ is chosen so that the minimal value of V on the border of $Base$ is greater than the maximal value of $\tilde{E}$ on $B^n(\mathbf{0}, 2r_1)$. On each geodesic line γ, connecting the south pole x_{sp} and the north pole x_{np}, define a cutoff function g_γ such that $g_\gamma(\vec{w}) = \tilde{E}(\gamma \cap B^n(\mathbf{0}, 2r_1))$ if $\varrho(x_{sp}, \vec{w}) \le 2r_1$ and $g_\gamma(\vec{w}) = V(\gamma \cap \partial Cap)$ if $\varrho(x_{sp}, \vec{w}) \ge R = 3r_1$. Define

$$\hat{E}(\vec{w}) := \begin{cases} \tilde{E}(\vec{w}) & \text{on int } B^n(\mathbf{0}, 2r_1), \\ g_\gamma(\vec{w}) & \text{on } Base \setminus \text{int } B^n(\mathbf{0}, 2r_1), \\ V(\vec{w}) & \text{on } Cap. \end{cases}$$

Thus we have obtained a potential $\hat{E} \in \mathcal{C}^2(\mathcal{M}_S^n)$ and, consequently, a dynamical system $(\mathcal{M}_S^n, \hat{\phi})$, generated by the gradient equation on $\mathcal{M}_S^n$

$$\dot{\vec{w}} = -\operatorname{grad} \hat{E}(\vec{w}) \tag{8}$$

has been obtained. By fixation of the time step and applying a Runge–Kutta method, cascades $(B^n(\mathbf{0}, 2r_1), \tilde{\phi}_T)$, $(B^n(\mathbf{0}, 2r_1), \tilde{\psi}_{\frac{T}{m}})$, $(\mathcal{M}_S^n, \hat{\phi}_T)$, and $(\mathcal{M}_S^n, \hat{\psi}_{\frac{T}{m}})$ are generated. By the fact shown above that $-\operatorname{grad} \tilde{E}(\vec{w})$ is nonzero on the border of the ball $B^n(\mathbf{0}, 2r_1)$ and points inward, the ball $B^n(\mathbf{0}, 2r_1)$ is an invariant set of the cascade $\tilde{\phi}_T$ and, for a sufficiently large m, of the cascade $\tilde{\psi}_{\frac{T}{m}}$ as well. This also implies invariance of $B^n(\mathbf{0}, 2r_1)$ for $\hat{\phi}_T$ and $\hat{\psi}_{\frac{T}{m}}$.

Step 3. Genericity As is known, Axiom A and the strong transversality condition are equivalent to structural stability of a dynamical system (see Palis and de Melo 1982, p. 171). On the other hand, for gradient dynamical systems, Axiom A implies that the system has only a finite number of singularities, all hyperbolic, whereas the strong transversality condition implies that the gradient system has no saddle–saddle connections. Thus, structural stability of the dynamical system $(\mathcal{M}_S^n, \hat{\phi})$, modeling a perceptron training process, implies the assumptions of Theorem 2.6. Moreover, the set of structurally stable systems is open and dense in the space of gradient dynamical systems $\mathcal{G}$ (see Palis and de Melo 1982, p. 116), which ensures that the properties specified in the assumptions of Theorem 2.6 are generic.

Step 4. Stability Under Numerics If a dynamical system generated by (2) has in the ball $B^n(\mathbf{0}, R) \subset Euc^n$ a finite number of singularities, all hyperbolic, then the dynamical system modeling a perceptron learning process (after compactification), generated by (8) on the manifold $\mathcal{M}_S^n$ satisfies the assumptions of Theorem 2.6 as well. This implies that Theorem 2.6 can be applied to the cascades $(\mathcal{M}_S^n, \hat{\phi}_T)$ and $(\mathcal{M}_S^n, \hat{\psi}_{\frac{T}{m},p})$ generated by (8). Thus, it is shown that a perceptron training process is, after compactification, generically stable under numerics with respect to the operator $\Psi = \hat{\psi}^m_{\frac{T}{m},p}$ according to every Runge–Kutta method $\hat{\psi}_{\frac{T}{m},p}$.

Step 5. Robustness In this step we want to prove 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. It will be shown that robustness is shared by learning processes resulting from vector fields belonging to some open and dense set in an appropriate space.

Lemma 3.2 *There exists an open and dense set of vector fields contained in $\mathcal{G}$ such that the cascade ϕ_T is $\mathcal{T}$ robust. Furthermore, for each $p \in \{1, 2, \dots\}$ and a sufficiently large m, the cascade $\Psi := \psi^m_{\frac{T}{n},p}$ is $\mathcal{T}$ robust as well, where $\psi_{h,p}$ is the diffeomorphism generated by a Runge–Kutta method of step size h and order p applied to the equation generating the flow ϕ.*

Proof Denote by $MS_{\mathcal{G}}$ the set of all Morse–Smale vector fields contained in $\mathcal{G}$ and recall that $\mathcal{G} \subset \Gamma$ is formed by all vector fields of the form $-\operatorname{grad} E$, where $E : \mathcal{M} \to \mathbb{R}$ is a C^2 function. The classical result is that the set $MS_{\mathcal{G}}$ is open and dense in $\mathcal{G}$; see for example Palis and de Melo (1982), p. 153.

On the other hand, if $-\operatorname{grad} E$ belongs to $MS_{\mathcal{G}}$, then the critical points of ϕ coincide with the fixed points of ϕ_T, and neither ϕ_T nor ϕ does admit other periodic orbits. Besides, stable and unstable manifolds of ϕ and ϕ_T at their (common) fixed points are the same. Hence, ϕ_T is a Morse–Smale diffeomorphism and, by Lemma 2.15, is $\mathcal{T}$ robust.

Also, one can easily see that for $-\operatorname{grad} E \in MS_{\mathcal{G}}$ all the assumptions of Theorem 2.6 are satisfied. Thus, ϕ_T and Ψ are topologically conjugate to each other if m is large enough; hence, by Theorem 2.14, we have also proved robustness of Ψ. This completes the proof of Lemma 3.2; consequently, the proof of Theorem 3.1 is completed as well. □

4 Practical Implications

The cascade Ψ describing a perceptron training process is a multi-step operator (note that the term *multi-step* should not be confused with a multi-step discretization method). This means that the m-fold iteration of the operator defined by a certain Runge–Kutta method is considered as a single unit of the theoretical analysis. It produces no limitations in practice since, during implementations, we can check the results of the training process after each m-step stage.

Theorems which were applied to the presented analysis describe certain properties of cascades on compact manifolds without boundaries. However, the numerical procedure, being the realization of the learning algorithm, is performed in $\mathbb{R}^n$. Therefore, we need conclusions concerning a set, say $\mathcal{A} \subset \mathbb{R}^n$, such that $B^n(\mathbf{0}, r_1) \subset \mathcal{A} \subset Base$ and $\alpha_m(\mathcal{A}) \subset Base$—see Steps 1 and 2 of the proof of Theorem 3.1.

Let us consider the set $\alpha_m(B^n(\mathbf{0}, 2r_1))$. We have $\alpha_m(B^n(\mathbf{0}, r_1)) \subset \alpha_m(B^n(\mathbf{0}, 2r_1))$ and, according to the fact that the conjugating homeomorphism α_m converges to identity for m converging to infinity (see Theorem 2.6), we also have $\alpha_m(B^n(\mathbf{0}, 2r_1)) \subset$ *Base* for a sufficiently large m. This means that topological conjugacy exists on the set $\mathcal{A} = B^n(\mathbf{0}, 2r_1)$. Thus, the considered cascades are also topologically conjugate on the subset of $\mathbb{R}^n$ on which the perceptron training process is implemented.

Although robustness is theoretically considered for $k \in \mathbb{Z}$, in implementations it is of interest for us only for $k \in \mathbb{N}$. The ball $B^n(\mathbf{0}, 2r_1)$ is positively invariant according to $\hat{\phi}_T$, and robustness is a topological conjugacy invariant (see Theorem 2.14). Therefore, according to the above conclusion concerning topological conjugacy, the robustness with respect to $\mathbb{N}$ takes place on the set $\mathcal{A}$ (see Remark 3).

Finally, we have to admit that the above result has a certain 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_T or $\psi_{\frac{T}{m},p}$ is inverse shadowing with respect to the class generated by real numerical methods like

$$\{\psi_{\frac{T}{m},p,b},\ b \in \{1, 2, \ldots\}\},$$

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

5 Concluding Remarks

In this paper we apply the fact that, on a certain n-dimensional manifold $\mathcal{M}_S^n$, homeomorphic to the sphere $\mathcal{S}^n$, a gradient dynamical system is, under some natural assumptions, correctly reproduced by its Runge–Kutta method of each order if only a single step of the numerical method is sufficiently small. The manifold $\mathcal{M}_S^n$ is constructed in such a way that the ball $B^n(\mathbf{0}, R)$ is a part of it. Therefore, the dynamical system $(B^n(\mathbf{0}, 2r_1), \check{\phi})$, modeling a perceptron learning process, remains unchanged after transforming of the problem onto the manifold in order to apply Theorem 2.6 for a perceptron learning process analysis. As the dynamics of gradient systems is very regular—in particular, the dynamics cannot be chaotic and there are no periodic orbits—Theorem 2.6 implies asymptotic stability of the learning process using every Runge–Kutta method, including the widely applied gradient descent method, which is simply the Euler method for (7). These properties are preserved under discretization and, due to the global topological conjugacy, when a Runge–Kutta method is applied. This implies $\mathcal{T}$ robustness of the learning process as well.

To sum up, the dynamics of learning processes of some artificial, nonlinear neural networks can be understood using dynamical systems theory, and in many situations the gradient dynamical systems are a good tool for that. It appears that, generically, for perceptrons such processes are convergent to equilibrium states and are both shadowing and inverse shadowing. This means that they are robust, and there may be good enough accuracy when they are performed by a computer. However, further studies on inverse shadowing with respect to a class of piecewise continuous methods are welcome.

References

Bielecki, A.: Dynamical properties of learning process of weakly nonlinear and nonlinear neurons. Nonlinear Anal., Real World Appl. **2**, 249–258 (2001)

Bielecki, A.: Estimation of the Euler method error on a Riemannian manifold. Commun. Numer. Methods Eng. **18**, 757–763 (2002)

Bielecki, A.: Topological conjugacy of discrete-time map and Euler discrete dynamical systems generated by gradient flow on a two-dimensional compact manifold. Nonlinear Anal., Theory Methods Appl. **51**, 1293–1317 (2002)

Bielecki, A., Ombach, J.: Shadowing property in analysis of neural networks dynamics. J. Comput. Appl. Math. **164–165**, 107–115 (2004)

Garay, B.: Discretization and Morse–Smale dynamical systems on planar discs. Acta Math. Univ. Comen. **63**, 25–38 (1994)

Hertz, J., Krogh, A., Palmer, R.G.: Introduction to the Theory of Neural Computation. Addison-Wesley, Redwood City (1991)

Hess, G.: Synaptic transmission and synaptic plasticity. In: Tadeusiewicz, R. (ed.) Theoretical Neurocybernetic. Warsaw University Press, Warsaw (2009) (in Polish)

Kloeden, P.E., Ombach, J.: Hyperbolic homeomorphisms are bishadowing. Ann. Pol. Math. **65**, 171–177 (1997)

Li, M.C.: Structural stability of Morse–Smale gradient-like flows under discretization. SIAM J. Math. Anal. **28**, 381–388 (1997)

Li, M.C.: Structural stability of the Euler method. SIAM J. Math. Anal. **30**, 747–755 (1999)

Lee, J.M.: Introduction to Smooth Manifolds. Springer, New York (2003)

Palis, J., de Melo, W.: Geometric Theory of Dynamical Systems. Springer, New York (1982)

Pilyugin, S.Yu.: Shadowing in Dynamical Systems. Lecture Notes in Mathematics, vol. 1706. Springer, Berlin (1999)

Rosenblatt, F.: Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington (1961)

Tadeusiewicz, R.: Problems of Biocybernetics. PWN, Warsaw (1994) (in Polish)

Wu, W., Xu, Y.: Deterministic convergence of an online gradient method for neural networks. J. Comput. Appl. Math. **144**, 335–347 (2002)