Preprint of Neural Computation article:2013 Mar;25(3):671-96. doi: 10.1162/NECO_a_00411. Epub 2012 Dec 28
-- this version is free from errors induced by NC typesetters --

Echo State Property Linked to an Input: Exploring a Fundamental Characteristic of Recurrent Neural Networks

G. Manjunath and H. Jaeger

School of Engineering and Science, Jacobs University Bremen gGmbH, 28759 Bremen, Germany.

Email: {m.gandhi, h.jaeger}@jacobs-university.de

Abstract

The echo state property is a key for the design and training of recurrent neural networks within the paradigm of reservoir computing. In intuitive terms this is a passivity condition: a network having this property, when driven by an input signal, will become entrained by the input and develop an internal response signal. This excited internal dynamics can be seen as a high-dimensional, nonlinear, unique transform of the input with a rich memory content. This view has implications for understanding neural dynamics beyond the field of reservoir computing. Available definitions and theorems concerning the echo state property, however, are of little practical use because they do not relate the network response to temporal or statistical properties of the driving input. Here we present a new definition of the echo state property which directly connects it to such properties. We derive a fundamental 0-1 law: if the input comes from an ergodic source, the network response has the echo state property with probability one or zero, independent of the given network. Furthermore we give a sufficient condition for the echo state property which connects statistical characteristics of the input to algebraic properties of the network connection matrix. The mathematical methods that we employ are freshly imported from the young field of nonautonomous dynamical systems theory. Since these methods are not yet well known in neural computation research, we introduce them in some detail. As a side story, we hope to demonstrate the eminent usefulness of these methods.

Keywords. Echo state property, Reservoir computing, Recurrent neural networks, Nonautonomous dynamical systems, Driven dynamical systems, Stability.

1 Introduction

In this article we derive a number of theoretical results concerning the dynamics of input-driven neural systems, using mathematical methods which still are widely unknown in neural computation and mathematical neuroscience. We hope that both – the results and the methods – will be of interest to the reader.

The *results* shed a new and sharp light on the question when an input-driven neural dynamics is "stable" or "unstable". These concepts are certainly understood in different ways in different communities and contexts. Here we address phenomena which are frequently described as "sensitive dependency on initial conditions" or "divergence of perturbed trajectories" or the like, and which are often related to "chaotic" dynamics. Such intuitions root in the theory of autonomous (i.e., not input-driven) dynamical systems. It is, in fact, not trivial to cleanly extend these intuitions to input-driven systems. We establish a rigorous formal framework in which *this* notion of stability becomes well-defined in input-driven systems, and prove a number of theorems. Among those, we derive a 0-1 law for systems driven by input from an ergodic source, to the effect that the driven system is "stable" with probability zero or with probability one.

Our work was originally motivated by questions which arise in the field of *reservoir computing* (RC), and more specifically, in the subfield of *echo state networks* (ESNs). ESNs are artificial recurrent neural networks (RNNs) which are used in machine learning for the supervised training of temporal pattern recognizers, pattern generators, predictors, controllers and more (short overview: (Jaeger, 2007); application oriented paper: (Jaeger et al., 2004); survey on state of the art: (Lukosevicius et al., 2009); other RC flavors besides ESNs: *liquid state machines* (Maass et al., 2002), *backpropagation-decorrelation learning* (Steil, 2004), *temporal recurrent neural network* (Dominey et al., 1995)). The basic idea behind RC is to drive a randomly created RNN (the *reservoir*) with the task input signal, and from the input-excited RNN-internal dynamics distil a desired output signal by a trainable readout mechanism – often just a linear readout trained by linear regression of the target output on the excited internal activation traces. A necessary enabling condition for this scheme to work is that the reservoir possesses the *echo state property* (ESP). This is a particular stability concept for which a number of equivalent definitions are available (Jaeger, 2001). Intuitively, these amount to the property that the reservoir dynamics asymptotically "washes out" initial conditions, or in other wordings, is "input-forgetting" or "state-forgetting". The ESP is connected to spectral properties of the network weight matrix, and some work has been spent on stating and refining these conditions (Jaeger, 2001; Buehner et al., 2006; Yildiz et al., 2012).

Importantly, the ESP is intrinsically tied to the characteristics of the driving input.

It may well be the case that for inputs of some kind, a reservoir does not forget initial states, while for others it does. Therefore, the ESP is not a property of a reservoir per se, but a property of a pair (reservoir, "set of admissible inputs"). Concretely, in all available definitions and conditions relating to the ESP, the admissible inputs are characterized solely by their value range. It is presupposed that the input takes values in a compact set U, from which the ESP becomes a property of a pair (reservoir, U). This setting has been the only one accessible to a mathematical treatment so far, which is why it is still there; but it is hardly relevant for the daily practice of reservoir computing and has given rise to widespread misconceptions (discussion in (Yildiz et al., 2012)).

The troublesome issue about specifying admissible inputs solely through their range is the following. Consider a standard discrete-time reservoir RNN with a tanh sigmoid. It is intuitively clear that the tanh mapping is more "contractive" for larger-amplitude neural activations than it is for small-amplitude ones, because the slope of the tanh is greatest around zero – larger arguments become more strongly quenched by the tanh tails. Thus, when a tanh reservoir is driven by large-amplitude input, the reservoir neurons will become highly excited, the tanh quenches strongly which results in an overall initial condition forgetting. In contrast, for small-amplitude input one may witness that the "washing out" characteristics becomes lost. In particular, a constant zero input is the most dangerous one for losing the ESP. But, very often in a practical application the relevant input range contains zero. One then earns $0 \in U$, and since all that is stated about possible inputs is their range, one also earns the constant-zero signal as an admissible input, which has to be accommodated into ascertaining the ESP. This is, firstly, unrealistic because more often than not in an application one will never encounter the constant-zero input. And secondly, this leads to unnecessarily strict constraints on the reservoir weight matrix because the ESP has to be also guaranteed for the zero input signal.

This situation has led to some confusion. On the one hand, in many published RC studies one finds an initial discussion on how the weight matrix was scaled to ensure the ESP (then the weight matrix is typically suboptimally scaled); on the other hand, in other studies one finds informal statements to the effect that a weight matrix scaling was used which formally aborts the ESP but was working well nonetheless because the input signal was strong enough (then there is no theoretical foundation for what was done). In some other published work the confusion culminates in the (incorrect) approach to scale the reservoir weight matrix "to the border of chaos" by setting it such that the ESP for zero input just gets (or gets not) lost (discussion of these issues in (Yildiz et al., 2012)).

All in all, an alternative definition of the ESP would be very welcome, which respects the nature of the expected input signals in more detail than just through fixing their range. We here provide such an alternative definition. In fact, we define the ESP

for a specific, *single* input signal $\{u_n\}_{n\in\mathbb{Z}}$. Our definition is not constrained to RNN settings but covers general input-driven dynamical systems provided their state space is compact. From this single-input-signal based definition of the ESP, we are able to derive the general 0-1 law mentioned previously (which boils down to the fact that if the ESP is obtained for a particular input signal, then with probability 1 it is also obtained for other inputs from the same source). Furthermore, returning to the specific case of tanh reservoirs, we relate the statistics of the input signal to spectral properties of the weight matrix, such that the ESP is guaranteed. While the bounds that we were able to spell out are still far from tight, we perceive these results as door-openers to further progress.

The *methods* which we use come from the young and strongly developing theory of *nonautonomous dynamical systems* (NDS). In mathematics, a (discrete-time) NDS is a dynamical system whose update map is time-varying. That is, while an autonomous dynamical systems is governed by a single update map $g : X \to X$, an NDS is updated by a different map g_n at every time $n \in \mathbb{Z}$ via $g_n : X \to X$. Input-driven systems are a special case of NDS: given a particular input sequence $\{u_n\}$, and an input-respecting update map $g : U \times X \to X$, one obtains the time-dependent maps g_n by $g_n(x) := g(u_n, x)$.

Biological and artificial neural information processing systems are almost always input-driven. The natural background theory to analyse them would thus be the theory of NDS. However, the theory of NDS is significantly more complex, significantly less developed, and much less known than the familiar theory of autonomous systems. It is also a "strange" world where familiar concepts like attractors and bifurcations reappear in new shapes and bear properties which are thoroughly different from the characteristics of their autonomous counterparts. Furthermore, a number of different basic concepts of attractivity are being used in the field. We have started to accommodate attractor concepts from NDS theory for neural dynamics phenomena elsewhere (Manjunath et al., 2012). Here we re-use some of the definitions and results from that work. For the purpose at hand, only quite elementary concepts from NDS theory are necessary. For readers not familiar with NDS, the present article might serve as a gentle first notice of the theory of NDS, and we hope that the benefits of this set of methods become clear. We provide essential references as our treatment below unfolds. Suggested readings containing important works in this area include (Kloeden et al., 2011) as a reference text, (Pötzsche, 2010) for a linearization theory, (Colonius et al., 2000) for nonautonomous control, (Rasmussen, 2007) for bifurcation theory and (Arnold, 1998) for random dynamics.

We hope that a wider usage of proper NDS concepts may help to make some of the neural dynamics analysis more rigorous, and also more appropriate to its subject, than it is possible when one only tries to adapt concepts to autonomous dynamics. In a short digression (Section 2.1) below we highlight the "hygienic" capacities of NDS

theory in a critique of "sensitivity to perturbation" analyzes which are sometimes found in the neural dynamics literature.

The article is organized as follows. We redraft the echo state property by defining it w.r.t. an input sequence in Section 2; we also analyze a simple special case where the input is periodic. Keeping in view that the ESP has a wider usage beyond artificial neural networks we spell out all our definitions for a general input driven dynamical system on a metric space. In Section 3 we prove a probability 0 or 1 determination of the echo state property for an input driven system. In Section 4 for a given artificial recurrent neural network with standard (tanh) sigmoid nonlinear activations, we establish sufficient conditions on the input for ESP to hold in terms of an induced norm of the internal weight matrix.

2 The Echo State Property w.r.t. an input

An input-driven system (IDS) on a metric space X is a continuous map $g : U \times X \to X$, where U is the metric space which contains the input driving sequence. In this paper we consider only those IDS for which X is compact and while U a complete metric space. This includes discrete-time recurrent neural networks whose neurons have bounded activations, and which are driven by $\mathbb{R}^K$-valued input signals. The dynamics of such an IDS, when driven by an input sequence $\{u_n\} \subset U$, is realized through $x_{n+1} = g(u_n, x_n)$. In the rest of this paper we denote an IDS by either $g : U \times X \to X$ or just by g with the assumptions of U being a complete and X a compact metric space implicit. Throughout we denote the diameter of a set $A \subset X$ by $\mathrm{diam}(A) = \sup\{d(x, y) : x, y \in X\}$. Also for a vector x in $\mathbb{R}^N$, we denote $\|x\|$ as its Euclidean norm and the operator or induced norm of any linear transformation T is denoted by $\|T\|$.

The state evolution in an IDS is studied through the orbits or solutions. Any sequence $\{\vartheta_n\} \subset X$ is called an *entire solution* of the IDS $g : U \times X \to X$ if there exists some $\{u_n\} \subset U$ such that $\vartheta_{n+1} = g(u_n, \vartheta_n)$ for all $n \in \mathbb{Z}$.

We now recall the original definition of the echo state property, which was stated for a recurrent neural network with a compact input space in (Jaeger, 2001). We formulate it in the more general framework of an IDS, and do not restrict the input space U to be compact.

Definition 2.1. (cf. (Jaeger, 2001).) Let $g : U \times X \to X$ be an input driven system, where X is compact and U is complete. A sequence $x_{(-\infty,0]} := (\ldots, x_{-1}, x_0) \subset X$ is said to be compatible with $u_{(-\infty,0)} := (\ldots, u_{-2}, u_{-1}) \subset U$ when $x_{k+1} = g(u_k, x_k)$ for

all $k < 0$. The input driven system $g : U \times X \to X$ has the *echo state property with respect to* U if for any given $u_{(-\infty,0)} \subset U$ and sequences $x_{(-\infty,0]}, y_{(-\infty,0]} \subset X$, both compatible with $u_{(-\infty,0)}$, the equality $x_0 = y_0$ holds.

A simple consequence of ESP w.r.t. input space in terms of entire solutions is the following:

Proposition 2.1. *Suppose $g : U \times X \to X$ is an input driven system which has the echo state property with respect to U. If $x_{(-\infty,0]}, y_{(-\infty,0]} \subset X$ are both compatible with $u_{(-\infty,0)}$ then $x_k = y_k$ for all $k \leq 0$. As a consequence, for any input sequence $\{u_n\}_{n\in\mathbb{Z}}$ there exists at most one entire solution.*

Proof. Since $x_{(-\infty,0]}$ and $y_{(-\infty,0]}$ are both compatible with $u_{(-\infty,0)}$ then by definition of compatibility in Definition 2.1 it follows that $x_{(-\infty,-1]}$ and $y_{(-\infty,-1]}$ are both compatible with $u_{(-\infty,-1)}$, where $x_{(-\infty,-1]} = (\ldots, x_{-2}, x_{-1})$, $y_{(-\infty,-1]} = (\ldots, y_{-2}, y_{-1})$ and $u_{(-\infty,-1)} = (\ldots, u_{-3}, u_{-2})$. Since g has ESP, by Definition 2.1, $x_{-1} = y_{-1}$. Repeating this argument, an obvious induction yields $x_k = y_k$ for all $k < 0$. Now if $x_0 = y_0$, then trivially by definition of an entire solution obtained from the input $\{u_n\}$ we have $x_k = y_k$ for all $k \geq 1$. Thus $x_k = y_k$ for all $k \in \mathbb{Z}$ and hence there exists at most one entire solution. ■

We now approach the core matter of this article, a treatment of the ESP at the resolution of individual input sequences.

Definition 2.2. An input driven system $g : U \times X \to X$ is said to have the *echo state property with respect to an input sequence* $\{u_n\}$ if there exists exactly one entire solution, i.e., if $\{\vartheta_n\}$ and $\{\theta_n\}$ are entire solutions, then $\vartheta_n = \theta_n$ for all $n \in \mathbb{Z}$.

This input-sequence sensitive definition of the ESP is related to the "classical" version, as follows. We will show below that any IDS $g : U \times X \to X$ has at least one entire solution for a given input sequence. Acknowledging this fact, it is then straightforward from Definition 2.1, Proposition 2.1 and Definition 2.2 that an IDS has the ESP w.r.t. the input space U if and only if it has the ESP w.r.t. every $\{u_n\} \subset U$.

For a deeper analysis of the ESP w.r.t. input sequences we will use methods from the theory of nonautonomous dynamical systems (e.g., (Kloeden et al., 2011)). Because these methods are not yet widely known in the neural computation world, we recall core concepts and properties.

A *discrete-time nonautonomous system* on a state space X is a (time-indexed) family of maps $\{g_n\}$, where each $g_n : X \to X$ is a continuous map, and the state of the

system, at time n, satisfies $x_n = g_{n-1}(x_{n-1})$. Since we will only be concerned with the discrete time case, we will drop the qualifier "discrete time" in the remainder of this text. Clearly an IDS gives rise to a nonautonomous system $\{g_n\}$ where $g_n(\cdot) := g(u_n, \cdot) : X \to X$. Following (Kloeden et al., 2011) and several other authors we recall the definition of what is called a *process* for a nonautonomous system. Although the term "process" has potentially confusing connotations, it is a standard terminology in the theory of nonautonomous dynamical systems, which makes us retain it. In essence, "process" here simply refers to a particular notation for a nonautonomous system, which will turn out to be very convenient:

Definition 2.3. Let $\mathbb{Z}^2_{\geq} := \{(n, m) : n, m \in \mathbb{Z} \,\&\, n \geq m\}$. A *process* ϕ on a state space X is a continuous mapping $\phi : \mathbb{Z}^2_{\geq} \times X \to X$ which satisfies the evolution properties:

(i). $\phi(m, m, x) = x$ for all $m \in \mathbb{Z}$ and $x \in X$.
(ii). $\phi(n, m, x) = \phi(n, k, \phi(k, m, x))$ for all $m, k, n \in \mathbb{Z}$ with $m \leq k \leq n$ and $x \in X$.

A sequence $\{\vartheta_n\} \subset X$ is said to be an *entire solution of* ϕ if $\phi(n, m, \vartheta_m) = \vartheta_n$ for all $n \geq m$.

It is readily observed that a nonautonomous system $\{g_n\}$ on X generates a process ϕ on X by setting $\phi(m, m, x) := x$ and $\phi(n, m, x) := g_{n-1} \circ \cdots \circ g_m(x)$. To verify that ϕ is a process we need to verify continuity. Continuity in the first two variables of ϕ is trivial. Also, the composition of finitely many continuous mappings makes the map $x_m \mapsto \phi(n, m, x_m)$ continuous, and hence ϕ is continuous. Conversely, for every given process ϕ on X, we obtain a nonautonomous system $\{g_n\}$ by defining $g_n(\cdot) := \phi(n + 1, n, \cdot)$. Likewise, the notion of an entire solution is equivalently transferred between the NDS and process formulation. Thus, a "process" and a "NDS" provide two equivalent views on the same object. We will switch between these views at our convenience.

Next, for each process ϕ on X we define a particular sequence of subsets of X, which carries much information about the qualitative behavior of the process:

Definition 2.4. Let ϕ be a process on a compact space X. The sequence $\{X_n\}$ defined by

$$X_n = \bigcap_{m<n} \phi(n, m, X)$$

is called the *natural association* of ϕ on X.

Since for any n

$$\phi(n + 1, n - 1, X) \subset \phi(n + 1, n, X), \tag{1}$$

X_n is a nested intersection of sets in Definition 2.4. It is clear that if some g_{n_0} in $\{g_n\}$ is not surjective on X, then the set $(g_{n_0}(X))^c$ (where $\cdot^c$ denotes set complement) is nonempty. Hence any entire solution of $\{g_n\}$ would not assume a value in the set $(g_{n_0}(X))^c$ at time n_0+1. Indeed a much stronger condition holds: X_n is exactly the set of points x through which some entire solution passes at time n. The natural association can thus be intuitively regarded as the (tight) "envelope" of all entire solutions. In order to ultimately establish this fact, we first note that the natural association is *ϕ-invariant*:

Proposition 2.2. (cf. (Manjunath et al., 2012)) *Let ϕ be a process on a compact space X. Then the natural association $\{X_n\}$ is such that each X_n is a nonempty closed subset of X and $\{X_n\}$ is ϕ-invariant, i.e., $\phi(n+1, n, X_n) = X_{n+1}$ for any $n \in \mathbb{Z}$ and hence for all $n \geq m$, $\phi(n, m, X_m) = X_n$.*

The proof is a straightforward exploitation of the compactness of X and can be found in (Manjunath et al., 2012) (also reproduced in Appendix A). Using this finding, one obtains the desired characterization of the natural association as "envelope" of entire solutions:

Lemma 2.1. (cf. (Manjunath et al., 2012)) *Let ϕ be a process on a compact space X. A sequence $\{X_n\}$ of subsets of X is the natural association of ϕ if and only if for all $k \in \mathbb{Z}$ it holds that*

$$X_k := \left\{ \pi_k(\vartheta_n) : \{\vartheta_n\} \text{ is an entire solution of } \phi \right\}, \tag{2}$$

where $\pi_k : \{\vartheta_n\} \mapsto \vartheta_k$ is the projection map.

Again, the proof is adapted from (Manjunath et al., 2012) and found in the Appendix A.

If g is an IDS, it follows from Proposition 2.2 and Lemma 2.1 that the set of entire solutions is nonempty. Hence in our definition 2.2 of the ESP w.r.t. $\{u_n\}$ the required existence of exactly one entire solution singles out the case of exactly one such solution against cases where there are more than one solutions. The case with no entire solution cannot arise.

We proceed to give a sufficient condition for a process to have exactly one entire solution. This condition is technical and will be used later in the proof of a core theorem.

Lemma 2.2. *Let ϕ be a process on a compact metric space X metrized by d. Suppose that, for all $n \in \mathbb{Z}$, there exists a sequence of positive reals $\{\delta_j\}_{j=1}^{\infty}$ converging to 0 such that $d(\phi(n, n-j, x), \phi(n, n-j, y)) \leq \delta_j \; d(x, y)$ for all $x, y \in X$ and for all $j \in \mathbb{N}$. Then there is exactly one entire solution of the process ϕ.*

Proof. Assume that there are two distinct entire solutions $\{\vartheta_n\}$ and $\{\theta_n\}$. Then $d(\vartheta_{n_0}, \theta_{n_0}) = \epsilon > 0$ for some n_0. For any such n_0, by hypothesis there exists $\{\delta_j\}_{j=1}^{\infty}$ converging to 0 such that

$$\begin{aligned} d(\phi(n_0, n_0 - j, \vartheta_{n_0-j}), \phi(n_0, n_0 - j, \theta_{n_0-j})) &\leq \delta_j \, d(\vartheta_{n_0-j}, \theta_{n_0-j}) \quad \forall j \\ &\leq \delta_j \, \mathrm{diam}(X) \quad \forall j. \end{aligned}$$

Since by definition of an entire solution, $\phi(n_0, n_0 - j, \vartheta_{n_0-j}) = \vartheta_{n_0}$ and $\phi(n_0, n_0 - j, \theta_{n_0-j}) = \theta_{n_0}$, we have $\epsilon < \delta_j \, \mathrm{diam}(X)$ for all j. But $\mathrm{diam}(X)$ is finite because X is compact, and hence $\delta_j \, \mathrm{diam}(X) \to 0$ as $j \to \infty$. This is a contradiction to the fact that $\epsilon > 0$. ■

Notice that in this lemma, the required null sequences $\{\delta_j\}_{j=1}^{\infty}$, which capture the rate of contraction "from the past to the present", depend on n. This allows for a time-varying degree of contractivity in the process. It is even possible that for limited periods, the maps $\{g_n\}$ are expanding. Specifically, consider the case of an input-driven recurrent neural network, with no internal weight adaptation (e.g., exploiting a network after a training phase). State contraction over time in the sense of the lemma is, in intuitive terms, related to input forgetting: when the contraction rate is high (i.e., $\{\delta_j\}_{j=1}^{\infty}$ converges quickly to zero), information about earlier input is quickly washed out from the network state. In a non-adapting RNN, the temporal variation of the contractivity of the process is entirely due to time-varying input itself. Again, still in purely intuitive terms, this means that some temporal input patterns can interact with the network such that they will be quickly forgotten, while other input patterns may be better preserved over longer timespans – or may, in fact, even become "enhanced" in the network state if the induced maps $\{g_n\}$ are expanding.

2.1 Some remarks on folklore

The ESP is constitutive for the Echo State Network (ESN) approach to training RNNs. In the concerned literature we witness that some assumptions are tacitly and pervasively taken for granted which actually have not been proven yet. Furthermore, we also witness some lack of conceptual rigor in some published work, especially with respect to the usage of notions from dynamical systems theory (attractors and chaos in particular). Here we want to clarify some of these themes and point to leads from nonautonomous dynamical systems theory.

First, we consider the case of an RNN which is driven by periodic input. This situation arises commonly when such systems are trained as periodic pattern generators or recognizers. It is taken for granted by ESN practitioners (the second author in-

cluded) that the induced network dynamics is (or asymptotically becomes) periodic of the same period. Our present setting helps us make this intuition rigorous.

Proposition 2.3. *Let $g : U \times X \to X$ be an input driven system with $\{u_n\}$ being p-periodic, i.e, the smallest positive integer s for which $u_{n+s} = u_n$ for all n is p. Suppose g has the ESP w.r.t. $\{u_n\}$, then the entire solution $\vartheta_n = \vartheta_{n+p}$ for all n. This entails that $\{\vartheta_n\}$ is r-periodic, with $p = kr$ for some integer k.*

Proof. Let ϕ be the process of the IDS corresponding to the input $\{u_n\}$ and $\{\vartheta_n\}$ be the entire solution of ϕ. From Lemma 2.1 we have $\vartheta_n = \bigcap_{m<n} \phi(n, m, X)$ for any n. Since $\{u_n\}$ is p-periodic, by definition of ϕ it follows that $\phi(n+p, m, X) = \phi(n, m, X)$ for all $m < n$. Hence $\vartheta_{n+p} = \bigcap_{m<n+p} \phi(n, m, X) \subset \bigcap_{m<n} \phi(n, m, X) = \vartheta_n$. Thus as subsets of X, the set inclusion $\vartheta_{n+p} \subset \vartheta_n$ holds. But since these sets are singletons and moreover nonempty, $\vartheta_{n+p} = \vartheta_n$ for any n. This directly entails that $\{\vartheta_n\}$ is r-periodic, with $p = kr$ for some integer k. ■

As a noteworthy special case, a constant (i.e. 1-periodic) input induces a constant entire solution $\vartheta_n \equiv \vartheta_0$. However, entire solutions extend from the infinite past, but in real life an RNN is started at some time 0 from some initial condition $x_0 \neq \vartheta_0$, after which its positive time evolution $\{g(u_n, x_n)\}_{n\geq 0}$ is observed. It follows easily from Lemma 2.1 and Proposition 2.3 that x_n converges to ϑ_0.

Jumping from the most simple (constant) to the most complex behavior, we would like to indulge in some cautionary remarks on using the notion of *chaos* when describing input-driven RNNs. We sometimes find in the literature discussions of RNNs driven by non-periodic input, where statements concerning chaotic behavior are made. Typically, chaos (or "edge of chaos") is identified numerically by simulation experiments. The RNN is driven several times with the same input sequence; at some time the network state is slightly perturbed and an exponential divergence of the perturbed trajectory from the unperturbed reference trajectory is quantified; if it is found positive then a chaotic dynamics is claimed.

This way of proceeding is dubious, and new approaches are needed for a number of reasons:

- The term *chaos* originates from the theory of autonomous systems – i.e., systems without input, or driven by constant or periodic input (in which cases they can be mathematically treated as autonomous). In the original context of autonomous systems, chaoticity is a property of *attractors* (sometimes more generally of *invariant sets*). Thus, for conceptual hygiene, it should be understood as a characteristic of attractors in nonautonomous systems as well.

But, in general, for nonautonomous systems attractivity notions are complicated, and typically an attractor is not a single subset of X but a particular sequence of subsets of X which is ϕ-invariant, where ϕ is the process of the nonautonomous system (Kloeden et al., 2011). The mathematical theory of attraction in nonautonomous dynamics is in its infancy, and a number of non-equivalent proposals for defining attractors have been forwarded, often depending on specific topological or probabilistic conditions. We bring to the attention of a reader interested in nonautonomous attractivity that the natural association $\{X_n\}$ of the process is also what is called a "pullback attractor" of the underlying nonautonomous system (Kloeden et al., 2011) (also see (Manjunath et al., 2012)). Further comments on attractor notions are beyond the scope of this paper.

- As we mentioned after Lemma 2.2, it may well be the case that for certain limited periods, a driving input leads to expanding mappings, while on longer timespans it will result in, on average, contracting dynamics. Specifically, if a standard RNN is driven with strong enough input, its units will be driven close to their saturation, which in turn leads to contractive dynamics in the sense of our Lemma 2.2 (or even stronger versions of contraction). If such contraction maps appear for longer timespans, the role of expanding maps diminishes or sensitivity on initial conditions may disappear eventually. Any numerical detection of "sensitivity on initial conditions" by perturbation experiments is only a temporally local finding in input driven systems and by itself cannot imply or preclude chaos. Thus, chaos has to be properly acknowledged to be an attribute of the input signal + driven system pair. In more technical terms, since chaos is an asymptotic concept (for instance, complexity quantifiers like Lyapunov exponents or topological entropy are obtained as time-related asymptotic quantities), to verify chaos in an input driven system the effect of the asymptotics of the input has to be factored into the picture. The notions of topological entropy for nonautonomous systems are under investigation and some interesting results can be found in (Kolyada et al., 1996; Oprocha et al., 2009; Zhang et al., 2009). Since an IDS $g : U \times X \to X$ and an input $\{u_n\}$ gives rise to a NDS $\{g_n\}$, calculating the topological entropy of $\{g_n\}$ actually quantifies the dynamical complexity by accounting for the input asymptotics. A formal calculation of topological entropy in this case is dependent on knowing the individual maps g_n explicitly unless algorithms are developed for estimating entropy from a time-series. When the input sequence $\{u_n\}$ is drawn from a finite-valued source, it seems that it is possible to estimate the lower bounds on the topological entropy by the methods developed in (Zhang et al., 2009). In a related finding, but in a very special case, an anonymous referee points us at (Amigó et al., 2012) where it is shown that when the input sequence $\{u_n\}$ is only an arbitrary switching between two values, and the two different g_n's obtained are affine transformations on the real line, then the topological entropy of the $\{u_n\}$ is identically equal to that of the

NDS it induces. However, this result does not hold when g_n's are not affine. For instance, in an artificial RNN by scaling the input sequence by a suitable large number the dynamics in the reservoir can be made trivial regardless of the input complexity. All of this prompts us to conclude that till estimates of topological entropy like in (Zhang et al., 2009) or other complexity measures factoring in the input asymptotics are obtained, no rigorous claims of chaos can be made based on perturbation-based detection experiments.

3 A 0 – 1 law for the ESP

In this section we consider an IDS $g : U \times X \to X$ with an input obtained as a realization of an U-valued stationary ergodic process $\{\xi_n\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$, i.e., for each ω and n, $\xi_n(\omega)$ takes values in the set U. Each realization $\{\xi_n(\omega)\}$, where $\omega \in \Omega$ gives rise to a separate nonautonomous system and hence has its own natural association $\{X_n(\omega)\}$. We thus consider $\{X_n\}$ to be a set-valued stochastic process. Before we embark on analyzing this object in more detail, we recall some standard notions from ergodic theory (e.g., (Billingsley, 1979; Krengel, 1985; Skorokhod et al., 2002; Walters , 1992).

Measure-theoretic dynamical systems and *measure preserving dynamical systems* (e.g., (Krengel, 1985; Walters , 1992)): A measure-theoretic dynamical system is a quadruplet $(\Omega, \mathcal{F}, \mu, T)$ where $(\Omega, \mathcal{F}, \mu)$ is a measure space and $T : \Omega \to \Omega$ is a measurable map. A measure-theoretic dynamical is said to be a measure preserving dynamical system (MPDS) if $\mu(T^{-1}(A)) = \mu(A)$ for all A in $\mathcal{F}$. A MPDS $(\Omega, \mathcal{F}, \mu, T)$ is said to be ergodic if for all $A \in \mathcal{F}$, $T^{-1}(A) = A$ implies $\mu(A) = 0$ or $\mu(A) = 1$.

Representing a stationary stochastic process as an MPDS (e.g., (Krengel, 1985)): Let $(\Omega, \mathcal{F}, P)$ be a probability space and S a separable complete metric space. Let $\mathcal{B}_S$ denote the Borel sigma-field of S. Let $\{\theta_n\}$ be an S-valued stationary process. Consider $(S^\infty, \mathcal{B}^\infty)$, where S^∞ is the Cartesian product of bi-infinite countable number of copies of S and $\mathcal{B}^\infty$ is the sigma-field generated by the product topology on S^∞. For each $\omega \in \Omega$, there exists an $\bar{u} = (\cdots, u_{-1}, u_0, u_1, \cdots) \in S^\infty$ such that $u_k = \theta_k(\omega)$. The process $\theta = \{\theta_n\}$ and P induce a measure μ on $(S^\infty, \mathcal{B}^\infty)$ defined by $\mu(A) := P(\theta^{-1}(A))$ for all $A \in \mathcal{B}^\infty$. It holds that the set $\{\bar{u} : \exists\, \omega \in \Omega \text{ such that } u_k = \theta_k(\omega)\, \forall\, k\}$ of all paths is in $\mathcal{B}^\infty$ and has μ-measure 1. The process $\{\theta_n\}$ is stationary if and only if $(S^\infty, \mathcal{B}^\infty, \mu, \sigma)$ is a MPDS where σ is the shift map that sends a point $\bar{u} := (\cdots, u_{n-1}, u_n, u_{n+1}, \cdots)$ in S^∞ to $\sigma(\bar{u}) = (\cdots, u_n, u_{n+1}, u_{n+2}, \cdots)$.

Ergodic stochastic processes (e.g., (Skorokhod et al., 2002, pp. 54–55)): An S-valued stationary process $\{\theta_n\}$ is said to be an ergodic process on $(\Omega, \mathcal{F}, P)$ if for every two

integers $l \leq m$ and for any finite collection of elements of the Borel sigma-field of S, $(A_l, A_{l+1}, \ldots, A_m)$ and $(B_l, B_{l+1}, \ldots, B_m)$, the limit

$$\lim_{j\to\infty} \frac{1}{j} \sum_{k=1}^{j} P(\{\theta_l \in A_l, \ldots, \ \theta_m \in A_m, \theta_k \in B_l, \cdots, \theta_{k+(m-l)} \in B_m\})$$

exists and is equal to

$$P(\{\theta_l \in A_l, \ldots, \theta_m \in A_m\}) P(\{\theta_l \in B_l, \ldots, \theta_m \in B_m\}).$$

Ergodic stochastic processes as ergodic MPDS (e.g., (Krengel, 1985)): It can be shown that $\{\theta_n\}$ is ergodic if and only if the MPDS $(S^\infty, \mathcal{B}^\infty, \mu, \sigma)$ is ergodic.

Birkhoff's ergodic theorem (e.g., (Krengel, 1985; Walters , 1992)): A core result in ergodic theory is Birkhoff's ergodic theorem. It states that if $\{\theta_n\}$ is an ergodic process and if $\Psi \in L^1(\mu)$ (i.e. Ψ is a complex valued function defined on S^∞ such that $\int |\Psi| dP_\mu < \infty$), then the limit $\lim_{j\to\infty} \frac{1}{j} \sum_{i=1}^{j} \Psi(\sigma^i(\theta(\omega)))$ exists μ-almost surely and when it exists, is equal to the μ-average:

$$\lim_{j\to\infty} \frac{1}{j} \sum_{i=1}^{j} \Psi(\sigma^i(\theta(\omega))) = \int \Psi \, d\mu \text{ almost surely w.r.t. } \mu. \tag{3}$$

A particular, useful application of (3) is if $\{\theta_n\}$ is an ergodic process and $\Phi \circ \theta_k$ belongs to $L^1(P)$ for some k (and hence for all k since for a stationary process $\int \Phi \circ \theta_k \, dP$ is independent of k), then

$$\lim_{j\to\infty} \frac{1}{j} \sum_{i=1}^{j} \Phi(\theta_i(\omega)) = \int \Phi \circ \theta_k \, dP \text{ almost surely w.r.t. } P. \tag{4}$$

Hausdorff semi-distance and Hausdorff metric: When X is a metric space with metric d, we denote by H_X the collection of all nonempty closed subsets of X. Let $dist(A, B) := \sup\{d(x, B) : x \in A\}$ be the *Hausdorff semi-distance* between any two $A, B \subset X$. It is well known that whenever X is complete (compact), H_X is also a complete (compact) metric space with the Hausdorff metric equivalently defined by $d_H(A, B) := \max(dist(A, B), dist(B, A)) := \inf\{\epsilon : A \subset B_\epsilon(B) \ \& \ B \subset B_\epsilon(A)\}$, where $B_\epsilon(A) := \{x \in X : d(x, A) < \epsilon\}$ is the open ϵ-neighborhood of A.

We now begin our analysis of an IDS whose input comes from a stationary ergodic source. The following lemma concerns the definition of a new real-valued process

obtained from an IDS when its input is $\{\xi_n\}$. Since notions of measurability are not obvious for set-valued functions, we prove measurability of the functions involved in Appendix A.

Lemma 3.1. *Consider an IDS $g : U \times X \to X$ and an U-valued ergodic process $\{\xi_n\}$ defined on $(\Omega, \mathcal{F}, P)$. For each realization $\{\xi_n(\omega)\}$, $\omega \in \Omega$, define the process $\phi_\omega(n, m, x) := g_{n-1,\omega} \circ \cdots \circ g_{m,\omega}(x)$, on X where $g_{n,\omega}(\cdot) := g(\xi_n(\omega), \cdot) : X \to X$. Let $\{X_n\}$ be the set-valued stochastic process where $\{X_n(\omega)\}$ is the natural association of the process ϕ_ω (notice that $X_n(\omega) \in \mathsf{H}_X$ as a consequence of Proposition 2.2). Define $\gamma : \mathsf{H}_X \to \mathbb{R}$ by*

$$\gamma(A) := \begin{cases} 0 & : \quad \text{if } diam(A) = 0, \\ 1 & : \quad \text{otherwise.} \end{cases}$$

Then $\{\gamma(X_n)\}$ is an ergodic stochastic process.

Proof. Recalling the definition of a natural association, we have $X_n(\omega) = \cap_{m<\infty}\phi_\omega(n, m, X)$. The set-valued function X_n is measurable by Lemma A.2. From Lemma A.3, we know $\gamma : \mathsf{H}_X \to \mathbb{R}$ is also a measurable function. Since the composition of two measurable functions is also measurable, $\gamma(X_n)$ is measurable for any n. Applying statement (ii) of Lemma A.1, we obtain $\{\gamma(X_n)\}$ to be an ergodic process. ■

We are now equipped for the main result of this section:

Theorem 3.1. *Let $\{\xi_n\}$ be an U-valued ergodic process defined on $(\Omega, \mathcal{F}, P)$ and $g : U \times X \to X$ an input driven system. Then the set of all $\omega \in \Omega$ such that g has the echo state property, i.e., the subset of Ω*

$$\left\{\omega : g \text{ satisfies ESP w.r.t. } \{\xi_n(\omega)\}\right\}$$

has either probability 1 or 0.

Proof. Let ϕ_ω, $\{X_n(\omega)\}$ and $\{\gamma(X_n)\}$ be defined as in Lemma 3.1. Since $X_n(\omega)$ is nonempty, $\text{diam}(X_n(\omega)) = 0$ if and only if $X_n(\omega)$ contains only a singleton of X. We also know from Lemma 2.1 that there is exactly one entire solution of ϕ_ω if and only if $\text{diam}(X_n(\omega)) = 0$ for all n. Using this and the definition of γ,

$$\left\{\omega : g \text{ satisfies ESP w.r.t. } \{\xi_n(\omega)\}\right\} = \left\{\omega : \gamma(X_n(\omega)) = 0 \ \forall n\right\}.$$

By Lemma 3.1 and Lemma A.1, $\{\gamma(X_{-n})\}$ is an ergodic process. Since γ takes the values 0 and 1, $\gamma \circ X_i$ belongs to $L^1(P)$ for any i. By Birkhoff's ergodic theorem in

(4) the limit

$$\lim_{j\to\infty} \frac{1}{j} \sum_{i=-1}^{-j} \gamma(X_i(\omega)) \tag{5}$$

exists and assumes the same value almost surely. We know a continuous map cannot map a singleton of X into a set of positive diameter. Hence, by definition of $\gamma(X_i(\omega))$ and Lemma 2.1 it follows that if $\gamma(X_i(\omega)) = 0$ for some i then $\gamma(X_m(\omega)) = 0$ for all $m \leq i$. Also by Lemma 2.1 it follows that if $\gamma(X_i(\omega)) = 1$ for some i then $\gamma(X_m(\omega)) = 1$ for all $m \leq i$. Hence the limit in (5) is equal to 0 if and only if $\gamma(X_i(\omega)) = 0$ for all i. Since the limit in (5) is almost surely the same constant, we have $\{\omega : \gamma(X_n(\omega)) = 0 \ \forall n\}$ has either probability 1 or 0. ■

4 Sufficient conditions for the ESP in an RNN

In this section we consider a discrete-time RNN with standard sigmoidal activations (with tanh nonlinearity) to provide sufficient conditions for the ESP w.r.t. an input. Sufficient conditions for the ESP w.r.t. an input space were provided by (Buehner et al., 2006; Yildiz et al., 2012) in terms of the internal weight matrix of the RNN. However, since our definition of the ESP is w.r.t. an input sequence, we have to bring in the role of the input as well into sufficient conditions for ESP. This is established in Theorem 4.1. When furthermore the input arises as a realization of a stationary ergodic source, we state sufficient conditions for the ESP w.r.t. typical realizations in Theorem 4.2. Since by its definition for a given IDS, the ESP w.r.t. an input depends upon the past history of the input, the higher order correlation or higher order statistics of the input data would be expected to play a role in determining the ESP. However, basing the ESP on higher order correlations or statistics of the input may not only be onerous but also of little help since complete higher order correlations or statistics are rarely available. In contrast, our sufficient conditions for the ESP in Theorem 4.1 and Theorem 4.2 are based on the intermittent frequencies of expanding and contracting behaviors of the nonautonomous system $\{g_n\}$ generated by the IDS and the input u_n via $g_n(\cdot) := g(u_n, \cdot)$.

Concretely, we consider the following standard RNN model, written as an IDS $g : \mathbb{R}^K \times \mathbb{R}^N \to \mathbb{R}^N$ given by

$$x_{n+1} = g(u_n, x_n) = \overline{\tanh}(W^{in}u_n + Wx_n), \tag{6}$$

where W^{in} and W are $K \times N$ and $N \times N$ dimensional real matrices representing the input and internal weight matrices of the neuronal connections, u_n and x_n are column vector representations, the function $\overline{\tanh} : \mathbb{R}^N \to (-1, 1)^n$ is defined by $\overline{\tanh} := (\tanh(y^1), \tanh(y^2), \ldots, \tanh(y^N))^\mathsf{T}$ when $y = (y^1, y^2, \ldots, y^N)^\mathsf{T}$ ($\cdot^\mathsf{T}$ denotes transpose).

Owing to the range of the tanh function, the effective dynamics of the IDS in (6) is always contained in $[-1,1]^N$. Thus, we can consider the IDS in (6) to be defined on $g : \mathbb{R}^K \times [-1,1]^N \to [-1,1]^N$. Note that we do not restrict the input space to be compact.

We make use of the following facts in proving Theorem 4.1. A generalization of the mean-value theorem in one-dimensional calculus in higher dimensions is the so-called mean-value inequality (e.g., (Furi et al., 1991)) and we state that if $\varphi : V \to \mathbb{R}^n$ is a C^1-function where V is an open subset of $\mathbb{R}^n$ then for any $x, y \in V$

$$\|\varphi(x) - \varphi(y))\| \leq \sup\{\|\varphi^{'}(z)\| : z \in V\}\|x - y\|, \tag{7}$$

where $\|\cdot\|$ is the Euclidean norm, and $\|\varphi^{'}(z)\| := \sup(\|\varphi^{'}(z)x\| : \|x\| = 1)$ is the induced norm of the Jacobian of $\varphi(\cdot)$ at the point z.

We also recall that $\tanh^{'}(x) = \text{sech}^{\,2}(x)$, where $\text{sech}\,(x) = \frac{2}{e^x + e^{-x}}$. Further $\text{sech}^{\,2}$ is such that

$$\begin{aligned}
\text{sech}^{\,2}(0) &= 1 && (8)\\
\text{sech}^{\,2}(x) &= \text{sech}^{\,2}(-x)\ \forall x \in \mathbb{R} && (9)\\
\text{sech}^{\,2}(|x|) &< \text{sech}^{\,2}(|y|) \text{ if } |y| < |x| && (10)\\
\text{sech}^{\,2}(|x|) \to 0 &\text{ as } |x| \to \infty && (11)
\end{aligned}$$

The following Lemma can be proved by elementary steps and is stated without proof.

Lemma 4.1. *Let $\{a_i\}_{i\in\mathbb{N}}$ be a sequence of real numbers such that $a_i > 0$ for all i. Then in the following statements the implications (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) hold:*

(i). $\limsup_{j\to\infty} \frac{1}{j} \sum_{i=1}^{j} \log(a_i) < 0$,

(ii). $\limsup_{j\to\infty} (\prod_{i=1}^{j} a_i)^{\frac{1}{j}} < 1$,

(iii). $\lim_{j\to\infty} \prod_{i=1}^{j} a_i = 0$.

Theorem 4.1. *Consider the IDS $g : \mathbb{R}^K \times [-1,1]^N \to [-1,1]^N$ defined in Equation (6) with an input $\{u_n\} \subset \mathbb{R}^K$. Then*

(i). g has the ESP w.r.t. $\{u_n\}$ if $\|W\| < 1$,

(ii). or in general i.e., even if $\|W\| \geq 1$, g has the ESP w.r.t. $\{u_n\}$ if $\{u_n\}$ is such that

$$\limsup_{j\to\infty} \frac{1}{j} \sum_{i=-1}^{-j} \Big(C_i - (1 + \ln(2))\Big) I\{C_i \geq 2\} > \frac{\ln(\|W\|)}{2}, \tag{12}$$

where C_i is the smallest absolute component of the vector $W^{in}u_i$: $C_i := \min(|W^{in}u_i|)$, and I is the indicator function, i.e., it takes $+1$ if its argument is true and 0 otherwise.

Proof. Let $X = [-1, 1]^N$. For any given sequence $\{u_i\}$, the IDS g specified in (6) defines a process ϕ given by $\phi(n, m, x) := g_{n-1} \circ \cdots \circ g_m(x)$, where $g_n(\cdot) := g(u_n, \cdot) : X \to X$. Also since g is C^1 continuous on the interior of $U \times X$ it follows that each g_n is C^1-continuous on the interior of X. Since a finite composition of C^1 functions is also C^1 continuous, the function $\phi(n, m, \cdot) : X \to X$ is C^1-continuous on the interior of X for any $n > m$.

Since $\phi(n, n-2, x) = \phi(n, n-1, \phi(n-1, n-2, x))$, by chain rule of differentiation, we know

$$\phi^{'}(n, n-2, x) = \phi^{'}(n, n-1, \phi(n-1, n-2, x))\phi^{'}(n-1, n-2, x).$$

In general for any $j \geq 1$,

$$\phi^{'}(n, n-j, x) = \prod_{k=0}^{j-1} \phi^{'}\Big(n-k, n-k-1, \phi(n-k-1, n-k-2, x)\Big). \tag{13}$$

Fix n. By applying the mean-value inequality (7), we get

$$d(\phi(n, n-j, x), \phi(n, n-j, y)) \leq \sup\{\|\phi^{'}(n, n-j, z)\| : z \in \mathrm{Int}(X)\}\, d(x, y), \tag{14}$$

where d is the Euclidean metric, $\|\phi^{'}(n, n-j, z)\|$ is the induced norm of the Jacobian of $\phi(n, n-j, \cdot)$ at the point z and Int(X) is the interior of X.

For any m, from (6) we know $\phi(m, m-1, x) = \overline{\tanh}(W^{in}u_{m-1} + Wx_{m-1})$. We know the derivative of the $\tanh(y^1)$ w.r.t. y^1 is $\mathrm{sech}^{\,2}(y^1)$. Define the function D_{u_m} by $x \in X \mapsto diag(\overline{\mathrm{sech}}^{\,2}(W^{in}u_m + Wx))$, where $diag(\star)$ denotes a $N \times N$ dimensional real valued diagonal matrix whose k-th diagonal element is $\star_k$, the k-th element of the vector $\star$, and the function $\overline{\mathrm{sech}}^{\,2} : y \mapsto (\mathrm{sech}^{\,2}(y^1), \mathrm{sech}^{\,2}(y^2), \ldots, \mathrm{sech}^{\,2}(y^N))^{\mathsf{T}}$. With this notation and by using the chain rule, $\phi^{'}(m, m-1, x) = D_{u_{m-1}}(x)W$. Again with such notation, from (13) and taking norms we can write

$$\|\phi^{'}(n, n-j, x)\| = \left\| \prod_{i=1}^{j} D_{u_{n-1}}\Big(\phi(n-i, n-i-1, x)\Big)W \right\|. \tag{15}$$

Proof of (i). We now proceed to find an upper bound on $\sup\{\|\phi^{'}(n, n-j, z)\| : z \in \mathrm{Int}(X)\}$. First, we find an upper bound on $\|D_{u_m}(\star)\|$ regardless of the argument $\star$.

We know that $D_{u_m}(\star)$ is a diagonal matrix, and hence $\|D_{u_m}(\star)\|$ is upper bounded by the maximum of the absolute value of the diagonal elements. Since the diagonal elements belong to the range of the sech^2 function, and sech^2 is a nonnegative function and can take a maximum value of 1, $\|D_{u_m}(\star)\| \leq 1$ for any u_m and any $\star$.

We can get a tighter upper bound on $\|D_{u_m}(\star)\|$ if u_m satisfies certain conditions. Denoting the maximum (minimum) of the elements of a vector v by $\max(v)$ ($\min(v)$),

$$\begin{aligned}
\|D_{u_m}(\star)\| &= \max(\overline{\text{sech}}^2(W^{in}u_m + \star)) \text{ by definition,} \\
&\overset{(10)}{=} \text{sech}^2(\min(W^{in}u_m + \star))) \\
&\overset{(9)}{=} \text{sech}^2(|min(W^{in}u_m + \star)|). \qquad (16)
\end{aligned}$$

Recall that $C_m = \min(|W^{in}u_m|)$, where $|\cdot|$ denotes the absolute value of its argument. Suppose $C_m \geq 2$ and let $\star$ take any value in $[-1, 1]^N$. Then by definition of C_m, clearly $|min(W^{in}u_m + \star)| \geq C_m - 1$. By (10), we have $\text{sech}^2(|min(W^{in}u_m + \star|) \leq \text{sech}^2(C_m - 1)$. Using this in (16) we can write

$$\|D_{u_m}(\star)\| \leq I\{C_m < 2\} + \text{sech}^2(C_m - 1)I\{C_m \geq 2\}. \qquad (17)$$

Let $\delta_j(n) := \sup\{\|\phi^{'}(n, n-j, z)\| : z \in \text{Int}(X)\}$. Using Lemma 2.2 in (14) we infer that if $\delta_j(n) \to 0$ as $j \to \infty$ for all n, then only one entire solution exists for ϕ and thus g has ESP w.r.t. $\{u_i\}$. We find an upper bound on $\delta_j(n)$ starting from the definition of $\|\phi^{'}(n, n-j, z)\|$ in (15):

$$\begin{aligned}
\delta_j(n) &= \sup_{x \in \text{Int}(X)} \left\| \prod_{i=1}^{j} D_{u_{n-i}}(\phi(n-i, n-i-1, x))W \right\| \\
&\overset{\|AB\| \leq \|A\|\|B\|}{\leq} \sup_{x \in \text{Int}(X)} \prod_{i=1}^{j} \left\| D_{u_{n-i}}\Big(\phi(n-i, n-i-1, x)\Big)W \right\|, \\
&\leq \|W\|^j \sup_{x \in \text{Int}(X)} \prod_{i=1}^{j} \left\| D_{u_{n-i}}\Big(\phi(n-i, n-i-1, x)\Big) \right\|, \\
&\overset{(17)}{\leq} \|W\|^j \prod_{i=1}^{j} I\{C_{n-i} < 2\} + \text{sech}^2(C_{n-i})I\{C_{n-i} \geq 2\} \qquad (18)
\end{aligned}$$

Clearly $\delta_j(n) \to 0$ as $j \to \infty$ whenever the right hand side of (18) converges to 0 as $j \to \infty$. The right hand side of (18) is a product of j positive reals and using

Lemma 4.1, $\delta_j(n) \to 0$ whenever:

$$\limsup_{j\to\infty} \frac{1}{j} \sum_{i=1}^{j} \ln(\|W\|^j \Big(I\{C_{n-i} < 2\} + \operatorname{sech}^2(C_{n-i} - 1) \Big) I\{C_{n-i} \geq 2\}) < 0,$$

$$\text{or if } \limsup_{j\to\infty} \frac{1}{j} \sum_{i=1}^{j} \ln \left(I\{C_{n-i} < 2\} + \operatorname{sech}^2(C_{n-i} - 1) I\{C_{n-i} \geq 2\} \right) < -\ln(\|W\|). \tag{19}$$

The left hand side of the inequality in (19) is upper bounded by zero since $\operatorname{sech}^2(\cdot)$ is upper bounded by 1. Let $\|W\| < 1$. Then the right hand side is positive (19) and hence the inequality in (19) is always true. Moreover (19) holds independent of any n, and hence whenever $\|W\| < 1$, $\lim_{j\to\infty} \delta_j(n) = 0$ for all n. Using Lemma 2.2, we infer that only one entire solution exists for ϕ and thus g has the ESP w.r.t. $\{u_i\}$.

Proof of (ii). We proceed to rearrange (19) and deduce further:

$$-\limsup_{j\to\infty} \frac{1}{j} \sum_{i=n-1}^{n-j} \ln \Big(\operatorname{sech}^2(C_i - 1) \Big) I\{C_i > 2\} > \ln(\|W\|), \tag{20}$$

$$\text{or } -\limsup_{j\to\infty} \frac{1}{j} \sum_{i=n-1}^{n-j} \ln \left(\frac{2}{e^{(C_i-1)} - e^{-(C_i-1)}} \right) I\{C_i > 2\} > \frac{\ln(\|W\|)}{2}, \tag{21}$$

$$\text{or } \limsup_{j\to\infty} \frac{1}{j} \sum_{i=n-1}^{n-j} \Big(-\ln 2 + (C_i - 1) + \ln(1 - e^{-2(C_i-1)}) \Big) I\{C_i > 2\} > \frac{\ln(\|W\|)}{2}, \tag{22}$$

$$\implies \limsup_{j\to\infty} \frac{1}{j} \sum_{i=n-1}^{n-j} \Big(C_i - (1 + \ln(2)) \Big) I\{C_i > 2\} > \frac{\ln(\|W\|)}{2}; \tag{23}$$

here (21) follows from (20) by definition of the function sech ; (22) follows from (21) by using $\ln(p/(q+r)) = \ln(p) - \ln(q) - \ln(1 - \frac{r}{q})$; (23) follows from (22) by using the fact that $\ln(1 - e^{-2(C_i-1)}) < 0$ (we can in fact ignore $\ln(1 - e^{-2(C_i-1)})$ without much loosening the bound as it attains values very close to zero whenever $C_i \geq 2$). If (23) holds for some n, by the property of lim sup, (23) also holds for any other n. Hence (12) is true if and only if (23) holds. Since (23) holds independent of any n, $\lim_{j\to\infty} \delta_j(n) \to 0$ for all n. This proves (ii). ■

The result from (ii) readily extends to a probabilistic version for networks driven by an ergodic source:

Theorem 4.2. *Let $\{\xi_n\}$ be an $\mathbb{R}^K$-valued ergodic process defined on $(\Omega, \mathcal{F}, P)$ such that the expectation of $\max(\xi_i)$ is finite, i.e., $-\infty < E(\max(\xi_i)) < \infty$, and let $g : \mathbb{R}^K \times [0,1]^N \to [0,1]^N$ as specified in Equation (6). Define the random variables $C_i(\omega) := \min(|W^{in}\xi_i(\omega)|)$ and*

$$\psi_i(\omega) := \begin{cases} C_i(\omega) & : \text{if } C_i(\omega) \geq 2, \\ 0 & : \text{otherwise.} \end{cases}$$

Then g has ESP almost surely whenever for some i the following inequality holds:

$$E[\psi_i] - (1 + \ln 2)P(\psi_i \geq 2) > \frac{\ln(\|W\|)}{2}.$$

Proof. Let $C_i(\omega) := \min(|W^{in}\xi_i(\omega)|)$. Then by (12) if

$$\limsup_{j\to\infty} \frac{1}{j} \sum_{i=-1}^{-j} \Big(C_i(\omega) - (1 + \ln(2)) \Big) I\{C_i(\omega) \geq 2\} > \frac{\ln(\|W\|)}{2} \tag{24}$$

holds, g has ESP w.r.t. $\{\xi_i(\omega)\}$.

Define random variables φ_i on the space Ω through $\varphi_i(\cdot) = \min \circ \text{abs} \circ W^{in} \circ \xi_i(\cdot)$, where abs stands for taking the absolute value of individual vector components. Furthermore, define random variables $\Psi_i(\cdot) := (\varphi_i(\cdot) - (1 + \ln 2))I\{\varphi_i(\cdot) > 2\}$. It can easily be verified that φ_i and hence Ψ_i belong to $L^1(P)$ if ξ_i is such that $|E(\max(\xi_i))| < \infty$. Now $(C_i(\omega) - (1 + \ln(2)))I\{C_i(\omega) \geq 2\} = \Psi_i(\omega)$. Hence we can apply Birkhoff's ergodic theorem (4) to evaluate the limit in (24).

In order to include the indicator function in the Lebesgue integral on the space Ω, for any $E \in \mathcal{F}$ define

$$\chi_E(*) := \begin{cases} 1 & : \text{ if } * \in E \\ 0 & : \text{ otherwise }, \end{cases}$$

Using this notation and applying (4), the limit in (24) exists and

$$\begin{aligned} &\lim_{j\to\infty} \frac{1}{j} \sum_{i=-1}^{-j} \Big(C_i(\omega) - (1 + \ln(2) \Big) I\{C_i(\omega) \geq 2\} \\ &= \int \Big(\min(|W^{in}\xi_i|) - (1 + \ln(2)) \Big) \chi_{\min(|W^{in}\xi_i|)\geq 2} \, dP \text{ a.s. w.r.t. } P, \\ &= E[\psi] - (1 + \ln 2)P(\{\omega : \psi_i \geq 2\}) \text{ a.s. w.r.t. } P. \end{aligned} \tag{25}$$

From this and (24), if $E[\psi] - (1 + \ln 2)P(\psi_i \geq 2) > \frac{\ln(\|W\|)}{2}$, we have the ESP w.r.t. almost all realizations of $\{\xi_i(\omega)\}$. Hence the theorem is proved. ■

The bounds offered by the theorems in this section are admittedly weak. Specifically, the reliance on $I\{C_i \geq 2\}$, with $C_i := \min(|W^{in}u_i|)$ will often bar practically useful applications of these theorems, since the condition $C_i \geq 2$ becomes easily unachievable when some of the input weights are small – which they usually are. However, we still think these results are worth reporting because they demonstrate the application of methods which may guide further investigations, eventually leading to tighter bounds.

5 Conclusion

With this article, we hope to have served two purposes. First, to put the field of reservoir computing on a more appropriate foundation than it had before, by presenting a version of the echo state property which respects inputs as individual signals – and not only as "anything that comes from some admissible value range". Second, to demonstrate the usefulness and power of concepts and insights from the field of nonautonomous dynamical systems.

A Appendix : Proofs and some intermediate results

Proof of Proposition 2.2. Since $\phi(n, m, \cdot)$ is continuous for every $n \geq m$ and X is compact, we have $\phi(n, m, X)$ is also a compact subset of X. Hence $\bigcap_{m<n} \phi(n, m, X)$ is an intersection of closed subsets of X which implies X_n is closed. Further it is also a nested intersection of closed sets in view of (1). Hence X_n is nonempty for any n.

We next prove the following **claim:** Let $A_1, A_2, \ldots,$ be a collection of nonempty subsets of X such that $A_{i+1} \subset A_i$ and Λ be a continuous function on X. Then

$$\Lambda(A) = \bigcap_{i=1}^{\infty} \Lambda(A_i), \text{ where } A := \bigcap_{i=1}^{\infty} A_i. \tag{A-26}$$

Proof of Claim. When $A := \bigcap_{i=1}^{\infty} A_i$, then one directly obtains $\Lambda(A) \subset \bigcap_{i=1}^{\infty} \Lambda(A_i)$.

Next, let us show the inclusion $\supset$. Let $y \in \bigcap_{i=1}^{\infty} \Lambda(A_i)$, i.e., $y \in \Lambda(A_i)$ for all $i \Longrightarrow$ there exists $x_i \in A_i$ such that $\Lambda(x_i) = y$. Let x' be an accumulation point of $\{x_i\}$. Since A is a nested intersection of A_i, by definition of lim sup of sets, $\limsup_{i\to\infty} A_i = \bigcap_{i=1}^{\infty} A_i = A$. Also by definition of lim sup of sets, all the accumulation points of $\{x_i\}$ are contained in A. Hence, $x' \in A$. By continuity of Λ, $\Lambda(x') = y$. Thus $y \in \Lambda(A)$.

Starting from the definition of X_n we deduce

$$\begin{aligned}
\phi(n+1, n, X_n) &= \phi(n+1, n, \bigcap_{m<n} \phi(n, m, X)) \\
&\overset{\text{(A-26)}}{=} \bigcap_{m<n} \phi(n+1, n, \phi(n, m, X)) \\
&= \bigcap_{m<n} \phi(n+1, m, X) \\
&\overset{(1)}{=} \bigcap_{m<n} \phi(n+1, m, X) \cap \phi(n+1, n, X) \\
&= \bigcap_{m<n+1} \phi(n+1, m, X), \\
&= X_{n+1}. \qquad \blacksquare
\end{aligned}$$

Proof of Lemma 2.1. We first show $\subset$ in (2). We first prove the following **claim:** A sequence of sets $\mathcal{A} = \{A_k\}$ is ϕ-invariant if and only if for every pair $k \in \mathbb{Z}$, $x \in A_k$ there exists an entire solution $\{\vartheta_n\}$ such that $\vartheta_k = x$ and $\vartheta_k \in A_k$ for all $k \in \mathbb{Z}$.

Proof of Claim. (from (Kloeden et al., 2011)) ($\Longrightarrow$) Let $k_0 \in \mathbb{Z}$ and choose $x \in A_{k_0}$. For $k \geq k_0$, define the sequence $\vartheta_k := \phi(k, k_0, x)$. Then by ϕ-invariance, $\vartheta_k \in A_k$ for any $k > k_0$. On the other hand, $A_{k_0} = \phi(k_0, k, A_k)$ for $k \leq k_0$ and so there exists a sequence $x_k \in A_k$ with $x = \phi(k_0, k, x_k)$ and $x_k = \phi(k, k-1, x_{k-1})$ for all $k \leq k_0$. Then define $\vartheta_k := x_k$ for $k \leq k_0$. This completes the definition of the entire solution ϑ_k.
($\Longleftarrow$) Suppose for any $k \in \mathbb{Z}$ and $x \in A_k$, there is an entire solution $\{\vartheta_n\}$ satisfying $\vartheta_k \in A_k$ for all $k \in \mathbb{Z}$. This implies $\phi(k+j, k, x) \subset A_{k+j}$ for all $j \geq 0$. Hence $\phi(k+j, k, A_k) \subset A_{k+j}$. The other inclusion follows from the fact that $\phi(k, k-j, \vartheta_{k-j}) = x$ for all $j \geq 0$.

Thus if $\{X_n\}$ is a natural association then it is ϕ-invariant by Proposition 2.2 and by the above claim $\subset$ in (2). We next show $\supset$ in (2). From the above claim, if there is an $x \in X_k$, then there is an entire solution $\{\vartheta_n\}$ such that $x = \vartheta_k$.

($\Longleftarrow$) Let $\{\vartheta_i\}$ be an entire solution. Now consider some ϑ_n. By definition there exists $x_k \in X$ such that $\phi(n, k, x_k) = \vartheta_n$ for all $k < n$. Clearly, $\phi(n, k, x_k) \in \phi(n, k, X)$

for all $k < n$. This implies $\vartheta_n \in \bigcap_{k<n} \phi(n, k, X) = X_n$. Since n was chosen arbitrarily, $\vartheta_i \in X_i$ for all i. ■

The following is an elementary result for stochastic processes (e.g., (Krengel, 1985)) which is recalled in our results:

Lemma A.1. *Let $\{\theta_n\}_{n\in\mathbb{Z}}$ be a S-valued ergodic process defined on $(\Omega, \mathcal{F}, P)$ Then*

(i). $\{\theta_{-n}\}$ is ergodic.
(ii). If R is some measurable space and $\Phi : S \to R$ is a measurable function and

$$\Theta_i : S \to R; \quad \Theta_i := \Phi \circ \theta_i,$$

then $\{\Theta_n\}$ is an R-valued ergodic process.

The following result is borrowed from another authors' manuscript (Manjunath et al., 2012) which is currently under review.

Lemma A.2. *Let the random variable X_n be defined as in Lemma 3.1. With respect to the given sigma algebra on Ω, and the Borel-sigma algebras defined on H_X obtained by the Hausdorff distance, $X_n : \Omega \to \mathsf{H}_X$ is a measurable function.*

Proof. For the U-valued stationary process $\{\xi_n\}$, let its MPDS be denoted by $(U^\infty, \mathcal{B}^\infty, \mu, \sigma)$.

Given any pair $i, n \in \mathbb{Z}$ such that $n > i$, we define $h_{n,i} : U^\infty \to \mathsf{H}_X$ by

$$h_{n,i}(\bar{u}) := g_{n-1} \circ \cdots \circ g_{k+1} \circ g_i(X),$$

where $\bar{u} = (\cdots, u_{-1}, u_0, u_1, \cdots) \in U^\infty$, $g_m : X \to X$ is defined by $g_m := g(u_m, \cdot)$, the map $g : U \times X \to X$ being as in Theorem 3.1.

Let d_U denote some metric on U that gives rise to $\mathcal{B}$. Then $d'_U := \min(1, d_U)$ also generates $\mathcal{B}$. Let $d_\infty(\bar{u}, \bar{v}) := \sum_{i=-\infty}^{\infty} \frac{d'_U(u_i, v_i)}{2^{|i|+1}}$ be the metric on U^∞. It may be verified that d_∞ generates the product topology on U^∞. We now claim that $h_{n,i} : U^\infty \to \mathsf{H}_X$ is a continuous map for each n and i. To show this let $\{\bar{u}_k\}$ be any sequence such that $\bar{u}_k \to \bar{u}$ as $k \to \infty$. We will show that $h_{n,i}$ is continuous by proving $h_{n,i}(\bar{u}_k) \to h_{n,i}(\bar{u})$ as $k \to \infty$.

Let $\bar{u}_k = (\cdots, u^k_{-1}, u^k_0, u^k_1, \cdots)$. Hence $h_{n,i}(\bar{u}_k) = g^{[k]}_{n-1} \circ \cdots \circ g^{[k]}_{k+1} \circ g^{[k]}_i(X)$, where $g^{[k]}_m := g(u^k_m, \cdot)$. Since $g : U \times X \to X$ is continuous, it follows from the continuity argument that given any ϵ there exists a $\delta > 0$ such that

$$\begin{aligned} d_U(u_m, u^k_m) < \delta \,\forall\, n \le m \le i \quad &\Rightarrow \quad d\big(g_{n-1} \circ \cdots \circ g_i(x)\ ,\ g^{[k]}_{n-1} \circ \cdots \circ g^{[k]}_i(x)\big) < \epsilon, \\ &\Rightarrow \quad h_{n,i}(\bar{u}_k) \subset B_\epsilon(h_{n,i}(\bar{u})) \ \ \& \ \ h_{n,i}(\bar{u}) \subset B_\epsilon(h_{n,i}(\bar{u}_k)), \\ &\Rightarrow \quad d_H(h_{n,i}(\bar{u}), h_{n,i}(\bar{u}_k)) < \epsilon. \end{aligned} \tag{A-27}$$

Since $\bar{u}_k \to \bar{u}$ as $k \to \infty$, we can find an integer K such that for all $k \ge K$, $d_\infty(\bar{u}, \bar{u}_k) < \frac{\delta}{2^{n-i}}$ holds. This implies $d_U(u_m, u^k_m) < \delta$ for all $n \le m \le i$. Hence for all $k \ge K$ from (A-27), we have $d_H(h_{n,i}(\bar{u}), h_{n,i}(\bar{u}_k)) < \epsilon$. Since ϵ was chosen arbitrarily, $h_{n,i}(\bar{u}_k) \to h_{n,i}(\bar{u})$ as $k \to \infty$. This implies that $h_{n,i}$ is continuous.

Define $h_n : U^\infty \to \mathsf{H}_X$ by

$$h_n(\bar{u}) := \bigcap_{j=1}^{\infty} h_{n,n-j}(\bar{u}). \tag{A-28}$$

Since $h_{n,n-j}$ is continuous for any $j \ge 1$, $h^{-1}_{n,n-j}(B) \in \mathcal{B}^\infty$ for any Borel subset B contained in H_X. This implies $h^{-1}_n(B) \in \mathcal{B}^\infty$ for any Borel subset B contained in H_X. This implies h_n is measurable.

For each $\omega \in \Omega$, there exists an $\bar{u}^\omega$ such that $u^\omega_k = \xi_k(\omega)$. Hence

$$\begin{aligned} X_n(\omega) \quad \overset{\text{by definition}}{=} \quad & \bigcap_{j=1}^{\infty} \phi_\omega(n, n-j, X), \\ = \quad & \bigcap_{j=1}^{\infty} g_{n-1} \circ \cdots \circ g_{n-j+1} \circ g_{n-j}(X), \ \text{where } g_j(\cdot) = g(\xi_j(\omega), \cdot), \\ = \quad & \bigcap_{j=1}^{\infty} h_{n,n-j}(\bar{u}^\omega), \\ = \quad & h_n(\bar{u}^\omega). \end{aligned}$$

Since $X_n(\omega) = h_n(\bar{u}^\omega)$, h_n is measurable, and $\omega \mapsto \bar{u}^\omega$ is obviously measurable, it follows that X_n is measurable. ■

Lemma A.3. *Let X be a compact metric space. Let a map $\gamma : \mathsf{H}_X \to \mathbb{R}$ be defined by*

$$\gamma(A) := \begin{cases} 0 & : \quad \textit{if } diam(A) = 0, \\ 1 & : \quad \textit{if otherwise.} \end{cases}$$

Then the function γ is measurable.

Proof. Since γ assumes values in $\{0, 1\}$, to prove γ is measurable it is sufficient to show that $\gamma^{-1}(0)$ is a Borel subset of H_X. We will show something stronger than

this by proving that $\gamma^{-1}(0)$ is a closed subset of H_X. Let $S_X := \{\{x\} : x \in X\}$, i.e., S_X is a set containing all singletons of the space X. By definition of γ, $\gamma(A) = 0$ if and only if A is a singleton of the space X. Hence $\gamma^{-1}(0) = S_X$. Now to show that the complement of S_X in H_X, i.e., $\mathsf{H}_X \setminus S_X$ is an open set we will prove that for any $e_X \in \mathsf{H}_X \setminus S_X$ there exists an open neighborhood of e_X contained in $\mathsf{H}_X \setminus S_X$. Let $e_X \in \mathsf{H}_X \setminus S_X$. Now considering e_X as a subset of the space X, we have at least two distinct elements $a, b \in X$ such that $a, b \in e_X$. Let $d(a, b) = 2\delta$ for some $\delta > 0$. By triangle inequality for any $x \in X$, it follows that at least one of the following holds: $d(a, x) \geq \delta$ or $d(b, x) \geq \delta$. Now if an open ball of radius δ around e_X does not intersect S_X if and only if $d_H(e_X, S_X) \geq \delta$:

$$\begin{aligned}
d_H(e_X, S_X) &:= \inf\Big(d_H(e_X, \{x\}) : \{x\} \in S_X\Big), \\
&= \inf\Big(\max(dist(e_X, \{x\}), dist(\{x\}, e_X)) : \{x\} \in S_X\Big), \\
&\geq \inf\Big(dist(e_X, \{x\}) : \{x\} \in S_X\Big), \\
&= \inf\Big(\sup(d(z, x) : z \in e_X) : \{x\} \in S_X\Big), \\
&\geq \inf\Big(\sup(d(a, x), d(b, x)) : \{x\} \in S_X\Big), \\
&\geq \delta \quad \text{(by triangle inequality).}
\end{aligned}$$

Thus the open ball of radius δ around e_X does not intersect S_X. Hence $\mathsf{H}_X \setminus S_X$ is open in H_X and hence S_X is closed and has to be a Borel subset of H_X. Thus γ is a measurable function. ■

Acknowledgements. The research reported here was funded by the FP7 European projects ORGANIC (`http://organic.elis.ugent.be/organic`) and AMARSi (`http://www.amarsi-project.eu/`).

References

Amigó, J.M., Gimenez, A., and Kloeden, P.E., (2012). Switching systems and entropy *submitted (preprint).*

Arnold, L., (1998). Random Dynamical Systems *Springer-Verlag, Heidelberg.*

Billingsley, P., (1979). Probability and Measure. *John Wiley, Chichester, England.*

Buehner, M., & Young, P., (2006). A tighter bound for the echo state property. *IEEE Trans. Neural Networks, 17(3)*, 820–842.

Colonius, F., & Kliemann, W., (2000). The Dynamics of Control. *Birkhauser*.

Dominey, P.F., Arbib, M., & Joseph, J.-P., (1995). A model of corticostriatal plasticity for learning oculomotor associations and sequences. *J. Cognitive Neuroscience*, *7*(3), 311 – 336.

Furi, M., & Martelli, M., (1991). On the Mean Value Theorem, Inequality and Inclusion. *Am. Math. Monthly*, *98*, 840 – 847.

Jaeger, H., (2001). The echo state approach to analysing and training recurrent neural networks. *Technical Report, GMD Report 148, GMD - German National Research Institute for Computer Science*, `http://minds.jacobs-university.de/pubs`.

Jaeger, H., & Haas, H., (2004). Harnessing nonlinearity, predicting chaotic systems and saving energy in wireless communication. *Science*, *304*(5667), 78 – 80.

Jaeger, H., (2007). Echo state network. *Scholarpedia*, *2*(9):2330.

Kloeden, P.E., & Rasmussen, M., (2011). Nonautonomous Dynamical Systems. *AMS Providence*.

Kolyada, S., & Snoha, L., (1996).. Topological entropy of nonautonomous dynamical systems. *Random Comput. Dynamics*, *4*(2-3), 205 – 233.

Krengel, U., (1985). Ergodic Theorems. *Walter de Gruyter, Berlin; New York*.

Lukoševičius, M., & Jaeger, H., (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, *3*(3), 127 – 149.

Maass, W., Natschläger, T., & Markram, H., (2002). Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Computation*, *14*(11), 2531 – 2560.

Manjunath, G. & Jaeger, H., (2012). The dynamics of random difference equations is remodeled by closed relations. *submitted (preprint)*, March.

Oprocha, P., & Wilczynski, P., (2009). Chaos in nonautonomous dynamical systems. *An. Stiint. Univ. Ovidius Constanta Ser. Mat.*, *17*(3), 209 – 221.

Pötzsche, C., (2010). Chaos in nonautonomous dynamical systems. *Springer Lecture Notes in Mathematics, Springer, Berlin*, *2002*.

Rasmussen, M., (2007). Attractivity and Bifurcation for Nonautonomous Dynamical Systems. *Springer Lecture Notes in Mathematics, Springer, Berlin, Heidelberg, New York*, *1907*.

Skorokhod, A.V., Hoppenstead, F.C., & Salehi, H., (2002). Random Perturbation Methods. *Springer Verlag, New York*.

Steil, J., (2004). Backpropagation-Decorrelation: online recurrent learning with O(N) complexity. *Proc. IJCNN*, *1*, 843 – 848.

Walters, P., (1992). An Introduction to Ergodic Theory. *Springer Verlag*.

Yildiz, I.B., Jaeger, H., & Kiebel, S.J., (2012). Re-visiting the echo state property. *Neural Networks*, DOI:10.1016/j.bbr.2011.03.031.

Zhang, J., & Chen, L., (2009). Lower bounds of the topological entropy for nonautonomous dynamical systems. *Appl. Math. J. Chinese Univ. Ser.*, *24*(1), 76 – 82.