Neural Networks 35 (2012) 1–9

Contents lists available at SciVerse ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Re-visiting the echo state property

Izzet B. Yildiz [a,*], Herbert Jaeger [b,1], Stefan J. Kiebel [a,2]

[a] Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstrasse 1A, 04103 Leipzig, Germany
[b] Jacobs University, Campus Ring 1, 28759 Bremen, Germany

ARTICLE INFO

Article history:
Received 29 January 2012
Received in revised form 24 June 2012
Accepted 12 July 2012

Keywords:
Echo state network
Spectral radius
Bifurcation
Diagonally Schur stable
Lyapunov

ABSTRACT

An echo state network (ESN) consists of a large, randomly connected neural network, the *reservoir*, which is driven by an input signal and projects to output units. During training, only the connections from the reservoir to these output units are learned. A key requisite for output-only training is the *echo state property* (ESP), which means that the effect of initial conditions should vanish as time passes. In this paper, we use analytical examples to show that a widely used criterion for the ESP, the spectral radius of the weight matrix being smaller than unity, is not sufficient to satisfy the echo state property. We obtain these examples by investigating local bifurcation properties of the standard ESNs. Moreover, we provide new sufficient conditions for the echo state property of standard sigmoid and leaky integrator ESNs. We furthermore suggest an improved technical definition of the echo state property, and discuss what practicians should (and should not) observe when they optimize their reservoirs for specific tasks.

1. Introduction

Echo state networks (ESN) Jaeger (2001) and Jaeger and Haas (2004) provide an architecture and supervised learning principle for recurrent neural networks (RNNs). The main idea is (i) to drive a random, large, fixed recurrent neural network with the input signal, thereby inducing in each neuron within this "reservoir" network a nonlinear response signal, and (ii) combine a desired output signal by a trainable linear combination of all of these response signals. The internal weights of the underlying reservoir network are not changed by the learning; only the reservoir-to-output connections are trained.

This basic functional principle is shared with Liquid State Machines (LSM), which were developed independently from and simultaneously with ESNs by Maass, Natschläger, and Markram (2002). An earlier precursor is a biological neural learning mechanism investigated by Peter F. Dominey in the context of modeling sequence processing in mammalian brains (Dominey, 1995). Increasingly often, LSMs, ESNs and some other related methods are subsumed under the name of *reservoir computing* (introduction: Jaeger (2007), survey of current trends: Lukoševičius and Jaeger (2009)). Today, reservoir computing has established itself as one of the standard approaches to supervised RNN training.

A crucial, enabling precondition for ESN learning algorithms to function is that the underlying reservoir network possesses the *echo state property* (ESP). Roughly speaking, the ESP is a condition of asymptotic state convergence of the reservoir network, under the influence of driving input. The ESP is connected to algebraic properties of the reservoir weight matrix, *and* to properties of the driving input. It is a rather subtle mathematical concept. Often the ESP is violated if the spectral radius of the weight matrix exceeds unity. Conversely, under rather general conditions, the ESP is obtained most of the time when the spectral radius is smaller than unity. This combination of facts has led to a widespread misconception that all one has to observe in order to obtain the ESP is to scale the reservoir weight matrix to a spectral radius below unity. We witness that a significant fraction – even a majority – of "end-users" of reservoir computing fall prey to this misconception. In fact, neither does a spectral radius below unity generally ensure the ESP, nor does a spectral radius above unity generally destroy it. In numerous applications – depending on the nature of the driving input and on the nature of the desired readout signal – a spectral radius well above unity serves best. The widespread practice of scaling the spectral radius to below unity thus leads to an under-exploitation of the learning and modeling capacities of reservoirs.

Here we re-visit the ESP, with the general aim to illuminate this concept from several sides for the practical benefit of reservoir computing practice. Besides this didactic goal, the technical contribution of this article is twofold. First, after summarizing the

* Corresponding author. Tel.: +49 341 9940 2216; fax: +49 341 9940 2221.
E-mail addresses: yildiz@cbs.mpg.de (I.B. Yildiz), h.jaeger@jacobs-university.de (H. Jaeger), kiebel@cbs.mpg.de (S.J. Kiebel).
URL: http://www.cbs.mpg.de/~yildiz (I.B. Yildiz).
1 Tel.: +49 421 200 3215; fax: +49 421 200 493215.
2 Tel.: +49 341 9940 2435; fax: +49 341 9940 2221.

0893-6080/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2012.07.005

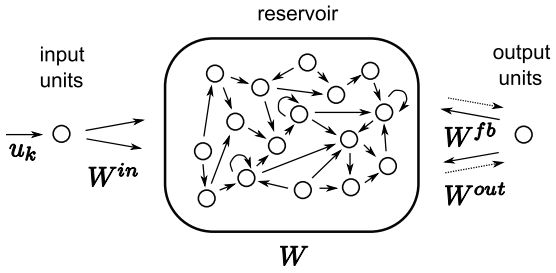

Fig. 1. The basic structure of an ESN. Solid arrows denote the fixed connections and dashed arrows denote the trainable connections.

standard formalism and ESP definition in Section 2, we present a bifurcation analysis to show in detail how the ESP can be lost even for spectral radii below unity (Section 3). Second, we derive a new, convenient-to-use formulation of a sufficient algebraic criterion for the ESP (Section 4). Then, in Section 5, we comment on situations where the ESP is obtained for spectral radii exceeding unity, which are of significant practical importance. We conclude with a short appreciation of the entire subject in a final discussion section.

2. Echo state networks

In this section we define the standard ESN and the echo state property.

The standard discrete-time ESN, which we denote shortly by $x_{k+1} = F(x_k, x_k^{out}, u_{k+1})$, is defined as follows:

$$\begin{aligned} x_{k+1} &= f(Wx_k + W^{in}u_{k+1} + W^{fb}x_k^{out}), \\ x_k^{out} &= g(W^{out}[x_k; u_k]), \end{aligned} \tag{1}$$

where $W \in \mathbb{R}^{N\times N}$ is the internal weight matrix or the reservoir, $W^{in} \in \mathbb{R}^{N\times K}$ is the input matrix, $W^{fb} \in \mathbb{R}^{N\times L}$ is the feedback matrix, $W^{out} \in \mathbb{R}^{L\times(N+K)}$ is the output matrix and $x_k \in \mathbb{R}^{N\times 1}$, $u_k \in \mathbb{R}^{K\times 1}$ and $x_k^{out} \in \mathbb{R}^{L\times 1}$ are the internal, input and output vectors at time k, respectively (see Fig. 1). The state activation function $f = (f_1, \ldots, f_N)^T$ is a sigmoid function (usually $f_i = \tan h$) applied component-wise with $f(0) = 0$ and the output activation function is $g = (g_1, \ldots, g_L)^T$ where each g_i is usually the identity or a sigmoid function. [;] denotes vector concatenation and x^T denotes the transpose of a vector x.

Here we consider only ESNs without feedback, i.e. $W^{fb} = 0$. The echo state network F with no feedback connection becomes:

$$x_{k+1} = F(x_k, u_{k+1}) = f(Wx_k + W^{in}u_{k+1}). \tag{2}$$

For the supervised learning algorithms which are used with ESNs (e.g. Jaeger (2001), Lukoševičius and Jaeger (2009)) it is crucial that the current network state x_k is uniquely determined by any left-infinite input sequence $\ldots, u_{k-1}, u_k$. This is made precise by requesting the *echo state property* (ESP). Since this effect depends on the input sequence, the definition of the ESP is stated relative to constraining the input range to a compact set U.

Concretely, we require the *compactness condition* which means F is defined on $X \times U$ where $X \subset \mathbb{R}^N$, $U \subset \mathbb{R}^K$ are compact sets and $F(x_k, u_{k+1}) \in X$ and $u_k \in U$, $\forall k \in \mathbb{Z}$. Note that the compactness of the state space X is automatically warranted when the reservoir unit nonlinearity f is bounded, like the $\tan h$ or the logistic sigmoid. Furthermore, in practical applications the input will always be bounded, so compactness of U will typically be warranted too.

Let $U^{-\infty} := \{u^{-\infty} = (\ldots, u_{-1}, u_0) \mid u_k \in U\ \forall k \le 0\}$ and $X^{-\infty} := \{x^{-\infty} = (\ldots, x_{-1}, x_0) \mid x_k \in X\ \forall k \le 0\}$ denote the set of left infinite input and state vector sequences, respectively. We say $x^{-\infty}$ is compatible with $u^{-\infty}$ when $x_k = F(x_{k-1}, u_k),\ \forall k \le 0$.

The definition of the echo state property when $W^{fb} = 0$ is as follows (adopted from Jaeger (2001)).

Definition 2.1 (*Echo State Property*)**.** A network $F : X \times U \to X$ (with the compactness condition) has the echo state property with respect to U: if for any left infinite input sequence $u^{-\infty} \in U^{-\infty}$ and any two state vector sequences $x^{-\infty}, y^{-\infty} \in X^{-\infty}$ compatible with $u^{-\infty}$, it holds that $x_0 = y_0$.

This "backward-oriented" definition can be equivalently stated in a forward direction. We remark that the original forward version given in Jaeger (2001) was too weak, and here present the corrected version (Jaeger, 2010). Similarly, let $U^{+\infty} := \{u^{+\infty} = (u_1, u_2, \ldots) \mid u_k \in U \forall k \ge 1\}$ and $X^{+\infty} := \{x^{+\infty} = (x_0, x_1, \ldots) \mid x_k \in X\ \forall k \ge 0\}$ denote the set of right-infinite input and state sequences, respectively.

Theorem 2.1 (*Forward Specification of ESP*)**.** *A network $F : X \times U \to X$ (with the compactness condition) satisfies the echo state property with respect to U if and only if it has the uniform state contraction property, i.e. if there exists a null sequence $(\delta_k)_{k\ge 0}$ such that for all $u^{+\infty} \in U^{+\infty}$, for all $x^{+\infty}, y^{+\infty} \in X^{+\infty}$ compatible with $u^{+\infty}$, it holds that for all $k \ge 0$, $\|x_k - y_k\| \le \delta_k$.*

For practical work with ESNs, it was mentioned in Jaeger (2001) that one usually obtains the echo state property by taking a random W and scaling it so that its spectral radius $\rho(W)$ is smaller than unity, where the spectral radius is the maximum of the absolute values of the eigenvalues of W. Although this recipe is used widely in reservoir computing practice, it is neither necessary nor sufficient to ensure the echo state property. In the next section, we investigate in more detail how and why it is not sufficient, and in Section 4 we provide a new, sufficient condition which is more practical than the current best known condition given in Buehner and Young (2006). In Section 5 we discuss relevant implications of the fact that it is not necessary, and comment on shortcomings of the current definition of the ESP.

3. Bifurcations in 2-dim echo state networks

Here we investigate ESNs with internal weight matrix W and a spectral radius $\rho(W) < 1$ where the network does not have the echo state property. We will constrain our analysis to the constant zero-input case, that is, $U = \{0\}$, because this basic case supports the present arguments already. In other words, we are interested in W matrices with $\rho(W) < 1$ for which the system

$$x_{k+1} = f(Wx_k) \tag{3}$$

does not have the echo state property. In particular, we investigate some bifurcation types which yield systems with non-trivial fixed points and periodic orbits. Note that the zero state (origin) is always a fixed point of the above system since $f(0) = 0$. For linear systems, $x_{k+1} = Wx_k$, the origin is indeed the global attractor of the system when $\rho(W) < 1$. The question is: For nonlinear systems, is the origin always the global attractor of the system in Eq. (3) when $\rho(W) < 1$? The answer is no and we give analytical examples below.

3.1. One dimensional case

When $x_k \in \mathbb{R}$ and $f = \tan h$, the fixed points can be found by solving $x = \tan h(wx)$ where $w \in \mathbb{R}$. In this case, the spectral radius is w. If $w > 1$, the origin is unstable and therefore we do not have the echo state property. If $w < 1$, then $\tan h(wx)$ is a contraction map and therefore the origin is globally asymptotically stable. Therefore, the examples we are interested in do not exist in one dimension.

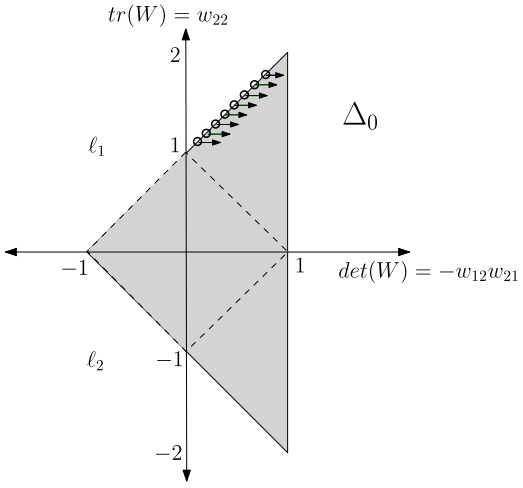

Fig. 2. The triangular region Δ_0 in the determinant-trace space. First, under some algebraic conditions described in the text, the system goes through a degenerate bifurcation along the side ℓ_1 of Δ_0 (shown by circles on ℓ_1). Then, pitchfork bifurcations occur towards the inside of Δ_0 which are shown by the horizontal arrows. The diamond-shaped region with dashed boundary gives the parameters for which W is diagonally Schur stable and therefore, no bifurcations can occur inside this region.

3.2. Two dimensional case

The analysis in two dimension is non-trivial. For a 2×2 weight matrix $W = \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix}$, the fixed points satisfy:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \tan h(w_{11}x_1 + w_{12}x_2) \\ \tan h(w_{21}x_1 + w_{22}x_2) \end{pmatrix}.$$

The first question is, for which values of w_{ij} does the matrix W have a spectral radius smaller than 1? This question was studied to investigate the region of stability for linear systems and is the well-known stability triangle (Thompson & Stewart, 2002):

$$\Delta = \{W \in \mathbb{R}^{2\times 2} : |tr(W)| - 1 < det(W) < 1\}$$

where $tr(W) = w_{11} + w_{22}$ is the trace of W and $det(W) = w_{11}w_{22} - w_{12}w_{21}$ is the determinant of W. One can show this by computing the eigenvalues of W directly and imposing the condition that both eigenvalues should have norm smaller than 1. Therefore, we are looking for W matrices in Δ for which the origin is not globally asymptotically stable.

Since the nonlinear system can be analyzed locally using the linear system and the origin is asymptotically stable in the linear case (in Δ), we can conclude that the origin is *locally* asymptotically stable in the nonlinear case. But there may be points away from the origin that are not attracted to the origin.

To be able to investigate this two dimensional system using one dimensional techniques, we also assume that $w_{11} = 0$. Then $tr(W) = w_{22}$ and $det(W) = -w_{12}w_{21}$. We call the corresponding stability triangle Δ_0 (see Fig. 2). The equation for fixed points reduces to:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \tan h(w_{12}x_2) \\ \tan h(w_{21}x_1 + w_{22}x_2) \end{pmatrix}. \tag{4}$$

Note that it is enough to consider $(x_1, x_2) \in (-1, 1) \times (-1, 1)$ since all states are mapped by $\tan h$ into this set after one iteration. The fixed points of this system can be found by solving $x_1 = \tan h(w_{12}x_2)$ and $x_2 = \tan h(w_{21}x_1 + w_{22}x_2)$. The second equation can be written as $\operatorname{arctanh}(x_2) = w_{21}x_1 + w_{22}x_2$. Plugging the first equation into the second one, we get $\operatorname{arctanh}(x_2) = w_{21}\tan h(w_{12}x_2) + w_{22}x_2$. Therefore, we need to find the fixed points of the function $\varphi(x_2)$ which for simplicity we write as $\varphi(x)$, where $\varphi(x)$ is:

$$\varphi(x) = -\frac{1}{w_{22}}(w_{21}\tan h(w_{12}x) - \operatorname{arctanh}(x)).$$

If this function, with W in the stability triangle, has a fixed point other than the origin, the echo state property does not hold.

Remark. Note that φ is not the system update equation for x_2, i.e. it does not hold that $(x_2)_{k+1} = \varphi((x_2)_k)$. The following bifurcation analysis of $\varphi(x)$ will only inform us about the creation of new fixed points for the system (4) as we move through parameter space within the stability triangle. The stability of these newborn fixed points has to be checked using local linearization (Jacobian) of the two dimensional system. Note that the origin is always going to be locally attracting (therefore stable) when the parameters are in Δ_0.

Remark. In Section 4.1, we will show that if W is diagonally Schur stable then the echo state property is satisfied for all inputs. In dimension 2, these matrices form the following set (Kaszkurewicz & Bhaya, 2000): $\{W \in \mathbb{R}^{2\times 2} : |det(W)| < 1, |w_{11} + w_{22}| < 1 + det(W)$ and $|w_{11} - w_{22}| < 1 - det(W)\}$. This corresponds to the diamond-shaped region in Δ_0 whose boundary is shown with dashed lines in Fig. 2. Therefore, bifurcations can only exist outside of this region.

First, we would like to show that along one of the sides of Δ_0, i.e. along $\ell_1 = \{w_{ij} | w_{11} = 0, w_{22} = -w_{12}w_{21} + 1$ and $0 \le w_{22} \le 2\}$, for a given *fixed* determinant and trace value (see little circles on the side ℓ_1 of Δ_0 in Fig. 2), it is possible to change w_{12} and w_{21} appropriately (keeping the determinant and trace constant) so that a degenerate bifurcation occurs. This bifurcation creates two more fixed points away from the origin. Moreover, following this degenerate bifurcation, if w_{12} and w_{21} are changed so that this time the determinant is increased and moved towards the inside of Δ_0 (horizontal arrows in the upper part of Fig. 2), then a pitchfork bifurcation occurs which creates two more fixed points from the origin. Therefore, we obtain examples which do not satisfy the echo state property even though $\rho(W) < 1$.

First, we look at the Taylor series expansion of $\varphi(x)$:

$$\begin{aligned} \varphi(x) &= -\frac{w_{21}}{w_{22}}\left(w_{12}x - \frac{(w_{12}x)^3}{3} + O(x^5)\right) \\ &\quad + \frac{1}{w_{22}}\left(x + \frac{x^3}{3} + O(x^5)\right) \\ &= -\frac{1}{w_{22}}(w_{12}w_{21} - 1)x + \frac{1}{3w_{22}}(1 + w_{12}^3 w_{21})x^3 + O(x^5) \\ &= px + qx^3 + O(x^5) \end{aligned} \tag{5}$$

where $p = -\frac{1}{w_{22}}(w_{12}w_{21} - 1) = (1 + det(W))/tr(W)$ and $q = \frac{1}{3w_{22}}(1 + w_{12}^3 w_{21})$. We investigate the following bifurcations.

3.2.1. Degenerate bifurcations

We start with bifurcations that occur on the line ℓ_1 at some fixed $(det(W), tr(W)) = (c, c + 1)$ value where $det(W) = c > 0$. Note that there are many w_{12} and w_{21} values which give the same determinant, i.e. $-w_{12}w_{21} = c$. Also note that p equals 1 on ℓ_1 and $q = \frac{1}{3w_{22}}(1 + w_{12}^3 w_{21})$.

We want to show that q changes sign as w_{12} and w_{21} are changed along the curve $-w_{12}w_{21} = c$ (see Fig. 3). Note that $q = 0$ if and only if $1 + w_{12}^3 w_{21} = 0$. Solving these two equations together, i.e. $-w_{12}w_{21} = c$ *and* $1 + w_{12}^3 w_{21} = 0$, one gets $w_{12} = \pm 1/\sqrt{c}$ and $w_{21} = \mp c\sqrt{c}$. In fact, when $|w_{12}| < 1/\sqrt{c}$ then $q > 0$ and when $|w_{12}| > 1/\sqrt{c}$ then $q < 0$. This means that the sign of q changes at this value.

To understand why the change of sign of q creates new fixed points, it is helpful to consider $y = \varphi(x) - x$ since the x-intercepts

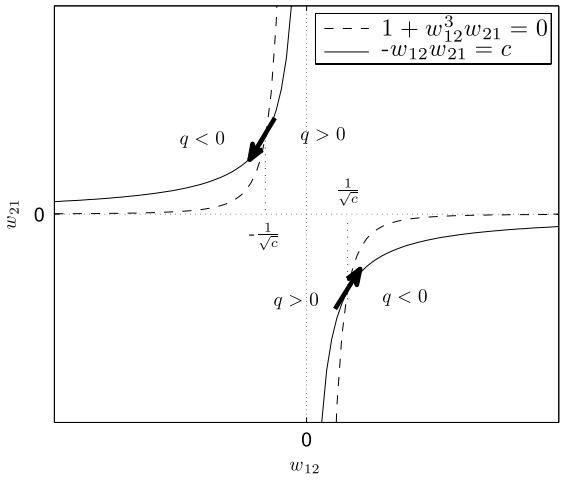

Fig. 3. The description of the degenerate bifurcation at a point $(det(W), tr(W)) = (c, c+1)$ where $0 < c < 1$ (shown with little circles on the side ℓ_1 of Δ_0 in Fig. 2). The degenerate bifurcation occurs as the sign of $q = \frac{1}{3w_{22}}(1+w_{12}^3 w_{21})$ changes from positive to negative. The parameters where this change of sign occurs are given by the curve $1+w_{12}^3 w_{21} = 0$. Since we also have the constraint $det(W) = -w_{12}w_{21} = c$, the bifurcations we describe in the text occur in the direction of the two black arrows.

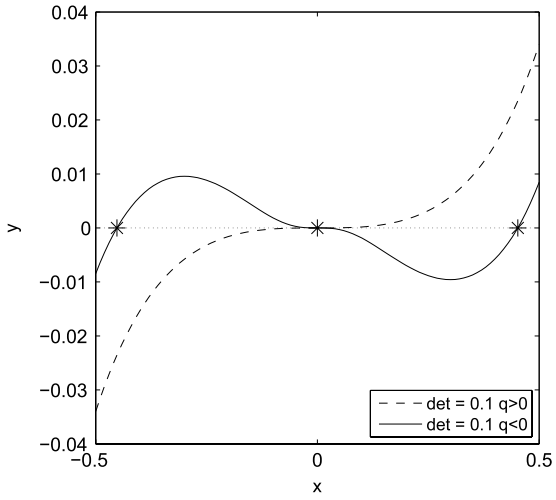

Fig. 4. The degenerate bifurcation occurs at the value $(det(W), tr(W)) = (0.1, 1.1)$ as q changes sign. Here, we draw $y = \varphi(x) - x$ for easier visualization of the fixed points which correspond to the x-intercepts of the graphs. Two new fixed points appear which are marked by asterisks. The entries used are $w_{11} = 0$, $w_{12} = 2$, $w_{21} = -0.05$ and $w_{22} = 1.1$ ($q \approx 0.18$) for the dashed curve and $w_{11} = 0$, $w_{12} = 10$, $w_{21} = -0.01$ and $w_{22} = 1.1$ ($q \approx -2.7$) for the solid curve. In both cases, $\rho(W) = 1$.

of this function gives the fixed points. Note that $y = \varphi(x) - x \approx qx^3$ around the origin and the sign of q changes this function locally (see Fig. 4). On the other hand, as $x \to 1^-$ and $x \to -1^+$, the arctanh(x) term in the definition of $\varphi(x)$ becomes dominant and therefore $y = \varphi(x) - x \approx \frac{1}{w_{22}}\text{arctanh}(x)$. This means, independent of the sign of q, $y = \varphi(x) - x$ approaches $+\infty$ and $-\infty$ as $x \to 1^-$ and $x \to -1^+$, respectively. Therefore, the local change in $y = \varphi(x) - x$ creates two new fixed points away from the origin as shown in Fig. 4. The existence of these fixed points can be shown rigorously using the Intermediate Value Theorem.

However, when $det(W) = c < 0$, i.e. $w_{12}w_{21} > 0$, we obtain $q = \frac{1}{3w_{22}}(1 + w_{12}^3 w_{21}) = \frac{1}{3w_{22}}1 + w_{12}^2(w_{12}w_{21}) > 0$ ($w_{22} = tr(W) > 0$ along ℓ_1). In other words, the sign of q does not change and no bifurcation occurs.

There are further bifurcations as the parameters are moved towards the *inside* of Δ_0.

3.2.2. Pitchfork bifurcation:

Eq. (5) is similar to the normal form of pitchfork bifurcation which is $px + qx^3$ where $q < 0$ for the supercritical case and $q > 0$ for the subcritical case (e.g. see Kuznetsov (2004)). In the supercritical case ($q < 0$), as p changes from $p < 1$ to $p > 1$, the fixed point at the origin changes its stability from stable to unstable, and two new, stable fixed points appear. In the subcritical case ($q > 0$), as p changes from $p > 1$ to $p < 1$, the fixed point at the origin changes its stability from unstable to stable, and two new, unstable fixed points appear.

We have shown above that for fixed values of $(det(W), tr(W))$ on ℓ_1, q can be positive or negative depending on w_{12} and w_{21}. Now, we look at what happens when $tr(W)$ is fixed but the determinant is increased so that we move from ℓ_1 towards the inside of Δ_0:

The case $det(W) > 0$ and $q < 0$.

When $det(W) = -w_{12}w_{21} = c > 0$, as mentioned in the degenerate bifurcation case, there exist w_{12} and w_{21} values such that $q < 0$. If we perturb w_{12} and w_{21} slightly, q stays negative since it is a continuous function. Therefore, one can indeed increase $det(W) = -w_{12}w_{21}$ slightly and preserve $q < 0$. As $det(W)$ increases, $p = -\frac{1}{w_{22}}(w_{12}w_{21} - 1)$ changes from $p = 1$ to $p > 1$ ($tr(W) = w_{22}$ is fixed but $det(W)$ increases towards the inside of Δ_0, see the horizontal arrows in the upper part of Fig. 2). Therefore, we observe a pitchfork bifurcation where the origin changes to

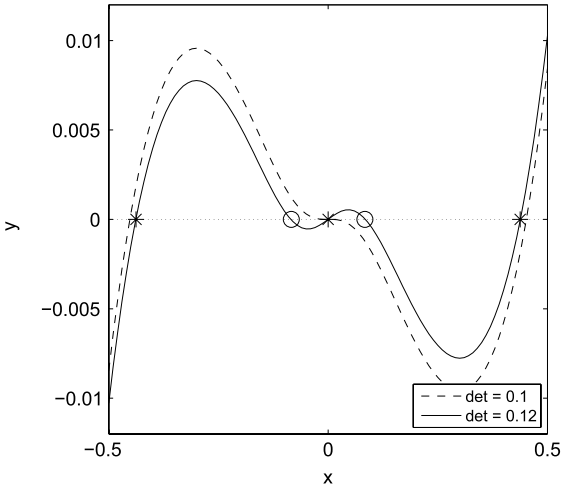

Fig. 5. The pitchfork bifurcation occurs along the side ℓ_1 of Δ_0 as the parameters are moved towards the inside of Δ_0 (see the horizontal arrows in the upper part of Fig. 2). Here, we draw $y = \varphi(x) - x$ for easier visualization of the fixed points which correspond to the x-intercepts of the graphs. Two new saddle fixed points appear which are marked by circles. The entries used are $w_{11} = 0$, $w_{12} = 10$, $w_{21} = -0.01$ and $w_{22} = 1.1$ for the dashed curve (which is the same as the solid curve in Fig. 4, $\rho(W) = 1$) and $w_{11} = 0$, $w_{12} = 10$, $w_{21} = -0.012$ and $w_{22} = 1.1$ for the solid curve ($\rho(W) = 0.9772$).

stable (since we move towards the inside of Δ_0) and two new fixed points appear in addition to the existing fixed points that appeared in the degenerate bifurcation (see Fig. 5). Thus, we now have five fixed points of φ. To further characterize the original two dimensional system, one can numerically compute the basin of attraction for each fixed point. Note that in Fig. 6, there are two attracting fixed points in addition to the origin itself. These are the fixed points that were born from the degenerate bifurcation described previously. By computing the Jacobians, one can see that that the new fixed points born from the pitchfork bifurcation correspond to saddle points in the original dynamics (4) (see Fig. 6). This gives an example of the case where $\rho(W) < 1$ and the echo state property does not hold. The case $det(W) > 0$ and $q > 0$.

When $det(W) > 0$, as described previously, there exist w_{12} and w_{21} values such that $q > 0$. As $p = -\frac{1}{w_{22}}(w_{12}w_{21} - 1)$ changes

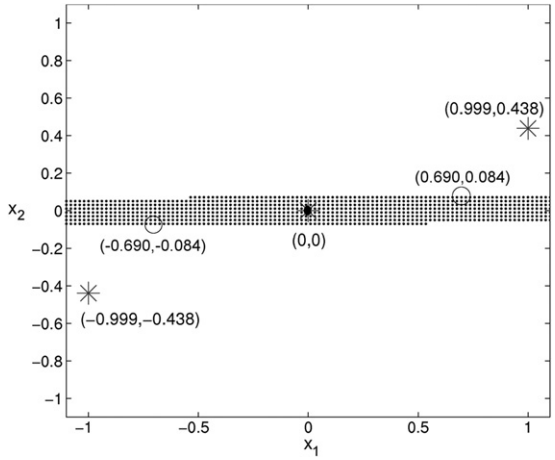

Fig. 6. The shaded region is the basin of attraction for the origin of the system with internal weight matrix $w_{11} = 0$, $w_{12} = 10$, $w_{21} = -0.012$ and $w_{22} = 1.1$. All the points (x_1, x_2) above this region are attracted to the stable fixed point (0.999, 0.438) and all the points (x_1, x_2) below this region are attracted to the fixed point $(-0.999, -0.438)$. Two saddle fixed points born from the pitchfork bifurcation are shown with circles and they lie on the boundary of the shaded region.

from $p > 1$ to $p < 1$, a pitchfork bifurcation occurs but this direction is towards the outside of Δ_0 and therefore not relevant here.

Remark. One can observe numerically that under some conditions, similar bifurcations can be observed along the side ℓ_2 of Δ_0 (Fig. 2) where the pitchfork bifurcation is replaced by period doubling bifurcation as the parameters are moved towards the inside of Δ_0. Since this requires the investigation of the second iteration of the system, we do not explore it analytically here.

3.2.3. Neimark–Sacker bifurcation

The Neimark–Sacker bifurcation is the discrete-time version of the Hopf bifurcation in the continuous case where the fixed point changes stability because of a pair of complex eigenvalues on the unit circle. At the critical parameter, the origin is surrounded by a closed invariant curve which can attract or repel nearby orbits. In dimension 2, we observed the Neimark–Sacker bifurcation only towards the outside of the stability triangle. However, one can find higher dimensional examples with $\rho(W) < 1$ where there exists an invariant orbit around the origin and therefore the echo state property is not satisfied. A numerically found, 4-dimensional example can be given with the internal weight matrix:

$$W = \begin{pmatrix} 0 & 1.95 & 3.3 & 1.56 \\ 0 & -0.47 & -1.38 & 1.3 \\ 0 & 0 & 0 & 1.95 \\ 0.23 & 0.67 & -1.13 & 0 \end{pmatrix}. \tag{6}$$

The eigenvalues of W are $0.2546 \pm 0.9201i$ and $-0.4896 \pm 0.5668i$ where the spectral radius is 0.9547. The invariant set which is the orbit of a high-order periodic point can be observed by taking arbitrary initial values and observing their forward iterations (Fig. 7).

3.3. Higher dimensions

The direct analysis of bifurcations in higher dimensions gets complicated quickly because of the number of parameters involved. However, one can always find examples in any dimension by generalizing from the lower dimensional examples. For example, let $W \in \mathbb{R}^{N\times N}$ be one of the interesting examples found above where the echo state property is not satisfied even though $\rho(W) < 1$. Then, let J_1, J_2 be two positive integers and consider the

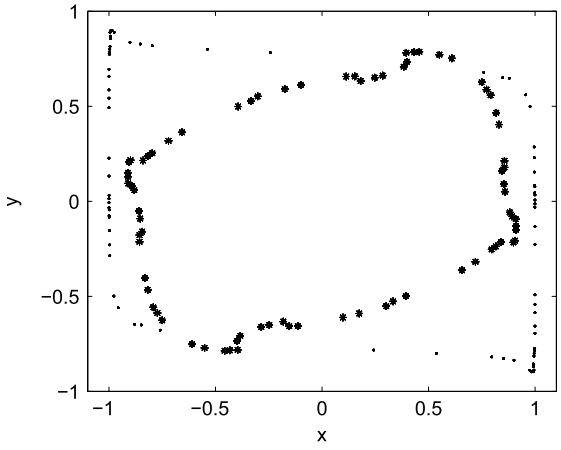

Fig. 7. Illustration of an example with $\rho(W) < 1$ for which the echo state property does not hold. The orbits of arbitrary initial points approach an invariant set where the internal weight matrix is given in Eq. (6). The orbit of an initial point (the first 8000 iterations are not shown) is attracted to a 4-dimensional high-order periodic orbit for which the dots show the projection to the first two coordinates and the asterisks show the projection to the last two coordinates.

matrix $\bar{W} = \begin{pmatrix} W & Q \\ 0 & R \end{pmatrix}$ where $Q \in \mathbb{R}^{N\times J_1}$ is an arbitrary matrix (no constraints) and $R \in \mathbb{R}^{J_2\times J_1}$ is an arbitrary matrix with $\rho(R) < 1$. This ensures that $\rho(\bar{W}) < 1$ and if $z \in \mathbb{R}^{N\times 1}$ is a fixed point of the system given by Eq. (3), then $\bar{z} = [z; 0] \in \mathbb{R}^{(N+J_1)\times 1}$ is a fixed point of the higher dimensional system with weight matrix $\bar{W}$.

In summary, using bifurcation analysis, we have shown that the condition $\rho(W) < 1$ is not a sufficient condition for the echo state property. The question remains how one can adapt this condition to establish the echo state property.

4. New sufficient conditions for the echo state property

In this section we provide sufficient conditions for the echo state property of the standard and leaky integrator ESNs. These sufficient conditions are important because, in practice, less restrictive conditions are typically used which do not guarantee the echo state property. In the standard ESNs (Eq. (1)), one usually samples a random internal weight matrix W with a subsequent scaling of the connectivity matrix W to ensure that its spectral radius is less than unity, $\rho(W) < 1$. In Section 3, we gave analytical examples for a simple but instructive case where the condition $\rho(W) < 1$ is insufficient to guarantee the echo state property. Similarly, for the leaky integrator ESNs, the commonly used condition, the *effective spectral radius* (see Section 4.2) being smaller than unity, is insufficient to obtain the echo state property as shown below in Example 4.1. Here, we derive new sufficient conditions and simple recipes to obtain the echo state property.

4.1. A sufficient condition for the standard ESN

A rather restrictive condition for the echo state property of the standard ESN was given in Jaeger (2001) as $\bar{\sigma}(W) < 1$ where $\bar{\sigma}(W)$ denotes the maximum singular value of W. Since this condition is too restrictive and the input is washed out very fast, it is not commonly used in practice. In this part, we state and prove Theorem 4.1 which provides a less restrictive condition in terms of diagonal Schur stability. We show that this condition is equivalent to the condition described in Buehner and Young (2006). The advantage of the Schur stability condition is that it is well studied in the literature. This enables us to list some important types of matrices which can be used as internal weight matrices guaranteed to have the echo state property. Finally, we give a simple recipe to obtain internal weight matrices which satisfy the echo state property.

We first define the following set of matrices which are important for the present analysis.

Definition 4.1. A matrix $W \in \mathbb{R}^{N\times N}$ is called *Schur stable* if there exists a positive definite symmetric matrix $P > 0$ such that $W^T PW - P$ is negative definite. If the matrix P can be chosen as a positive definite *diagonal* matrix, then W is called *diagonally Schur stable*. The positive definite and negative definite matrices are denoted by $P > 0$ and $P < 0$, respectively.

The notion of diagonal Schur stability is enough to state our result.

Theorem 4.1. *The network given by Eq. (2) with internal weight matrix W satisfies the echo state property for any input if W is diagonally Schur stable, i.e. there exists a diagonal $P > 0$ such that $W^T PW - P$ is negative definite.*

The proof of this theorem is given in the Appendix.

The diagonal Schur stability was investigated in Bhaya and Kaszkurewicz (1993) and more recently, in Kaszkurewicz and Bhaya (2000), the authors included matlab code for checking whether a given matrix is diagonally Schur stable. Moreover, the following set of matrices are proven to be diagonally Schur stable. Therefore, the echo state property for all inputs is satisfied for internal weight matrices which fulfill one of the following criteria:

- $W = (w_{ij})$ such that $\rho(|W|) < 1$ where $|W| = (|w_{ij}|)$.
- $W = (w_{ij})$ such that $w_{ij} \geq 0, \forall i, j$ and $\rho(W) < 1$.
- W such that $\rho(W) < 1$ and there exists a nonsingular diagonal D such that $D^{-1}WD$ is symmetric (this also includes symmetric matrices).
- W is a triangular matrix and $\rho(W) < 1$.
- $W \in \mathbb{R}^{2\times 2}$, $|det(W)| < 1$, $|w_{11} + w_{22}| < 1 + det(W)$ and $|w_{11} - w_{22}| < 1 - det(W)$.

More examples such as quasidominant and checkerboard matrices are given in Kaszkurewicz and Bhaya (2000) with the relevant definitions.

A simple recipe for the echo state property of the standard ESNs

The condition $\rho(|W|) < 1$ where $|W| = (|w_{ij}|)$ gives a simple way to construct internal weight matrices for the standard ESNs that satisfy the echo state property:

(i) Start with a random W with all non-negative entries, $w_{ij} \geq 0$.
(ii) Scale W so that $\rho(W) < 1$.
(iii) Change the signs of a desired number of entries of W to get negative connection weights as well.

Note that this recipe is more restrictive than the necessary condition $\rho(W) < 1$; however, the echo state property is guaranteed for any input.

Remark. The diagonal Schur stability condition turns out to be equivalent to the condition given in Buehner and Young (2006). The authors have shown that the echo state property is satisfied if $\inf_{D\in\mathcal{D}} \bar{\sigma}(DWD^{-1}) < 1$ where $\mathcal{D}$ is the set of nonsingular diagonal matrices and $\bar{\sigma}$ is the largest singular value. The equivalence of this condition to the diagonal Schur stability can be obtained by noting (Packard & Doyle, 1993): (i) The infimum does not change when we restrict $\mathcal{D}$ to be the set of *positive* diagonal matrices since any nonsingular diagonal matrix D can be written as $D = UD_+$ where U is a diagonal matrix with entries ± 1 (therefore unitary) and D_+ is a positive diagonal matrix. Therefore, $\inf_{D\in\mathcal{D}} \bar{\sigma}(DWD^{-1}) = \inf_{D_+\in\mathcal{D}} \bar{\sigma}(D_+WD_+^{-1})$. (ii) The following equivalences hold: $\bar{\sigma}(DWD^{-1}) < 1 \Leftrightarrow \rho(D^{-1}W^TDDWD^{-1}) < 1^2 = 1 \Leftrightarrow D^{-1}W^TDDWD^{-1} - I < 0 \Leftrightarrow W^TD^2W - D^2 < 0$ where the last inequality is equivalent to being diagonally Schur stable.

Next, we use similar techniques to find sufficient conditions for the echo state property in the leaky integrator case.

4.2. Sufficient conditions for leaky integrator ESNs

In this section, we will first describe the leaky integrator ESNs and state the discretized version. Then, in Theorem 4.2, we will define a new system (Eq. (8)) such that its stability implies the echo state property for the discretized leaky integrator ESN (Eq. (7)) with no feedback. There are various ways to check the stability of this new system and we give two closely related conditions as corollaries. One of these corollaries provides a simple recipe to obtain internal weight matrices with the echo state property.

For dealing with slowly and continuously changing systems, a continuous version of ESN was introduced in Jaeger (2001) and investigated in more detail in Jaeger, Lukosevicius, Popovici, and Siewert (2007). It is defined by:

$$\dot{x} = \frac{1}{c}(-ax + f(W^{in}u + Wx + W^{fb}x^{out})),$$
$$x^{out} = g(W^{out}[x; u]),$$

where the matrices $W^{in} \in \mathbb{R}^{N\times K}$, $W \in \mathbb{R}^{N\times N}$, $W^{out} \in \mathbb{R}^{L\times(K+N)}$ and $W^{fb} \in \mathbb{R}^{N\times L}$ are the input, internal, output and feedback connection weight matrices, respectively. $u \in \mathbb{R}^K$ is the external input, $x \in \mathbb{R}^N$ is the internal weight activation state, $x^{out} \in \mathbb{R}^L$ is the output vector, $c > 0$ is a global time constant, $a > 0$ is the leaking rate, f is a sigmoid function (usually $\tan h$ applied component-wise), g is the output activation function (usually the identity or a sigmoid) and $[;]$ denotes vector concatenation.

The Euler discretization with step-size $\delta > 0$ gives the following discrete-time version with a discrete-time sampled input u_k^δ. We use the notation $\Delta t = \frac{\delta}{c}$ for simplicity:

$$x_{k+1} = (1 - a\Delta t)x_k + \Delta t f(W^{in}u_{k+1}^\delta + Wx_k + W^{fb}x_k^{out}), \quad (7)$$
$$x_k^{out} = g(W^{out}[x_k; u_k^\delta]).$$

It is assumed that $a\Delta t < 1$. A sufficient condition for the echo state property of this discretized version with no feedback connection was given in Jaeger (2001): If $|1 - \Delta t(a - \sigma_{max})| < 1$ where σ_{max} is the largest singular value of W, then the echo state property is satisfied. Furthermore, it was also stated that if the matrix $\tilde{W} = \Delta tW + (1 - a\Delta t)I$ where I is the identity matrix, has a spectral radius $\rho(\tilde{W}) > 1$ then the echo state property (for zero input) is not satisfied since the origin becomes unstable. In practice, the necessary condition $\rho(\tilde{W}) < 1$ is used to obtain stable leaky integrator ESN's where $\rho(\tilde{W})$ is called the *effective spectral radius*. However, one should be aware that the necessary condition $\rho(\tilde{W}) < 1$ is indeed not sufficient and counterexamples can be given where the echo state property is not satisfied.

Example 4.1. Let $a = 1$, $c = 1$, $\delta = 0.1(\Delta t = 0.1)$ and $\tilde{W} = \Delta tW + (1 - a\Delta t)I = \begin{pmatrix} 0 & 10 \\ -0.012 & 1.1 \end{pmatrix}$. Note that the effective spectral radius, $\rho(\tilde{W}) = 0.9772 < 1$. However, the system given in Eq. (7) with zero-input, no feedback and internal weight matrix W, has fixed points other than the origin and therefore it does not satisfy the echo state property. In particular, the fixed points $(-0.999, -0.943)$ and $(0.999, 0.943)$ also attract nearby points. The phase space is very similar to the one given in Fig. 6 which was investigated for the standard ESNs in more detail in Section 3.

Since the commonly used effective spectral radius condition, $\rho(\tilde{W}) < 1$, is not sufficient for the echo state property, it is still important to find a less restrictive sufficient condition than the one given by $|1 - \Delta t(a - \sigma_{max})| < 1$. In fact, using a similar idea as in the proof of Theorem 4.1, we will define a new system (Eq. (8)) for which, once stability is established, the echo state property is implied for the corresponding leaky integrator ESN.

Theorem 4.2. *The leaky integrator ESN given in Eq. (7) with $W^{fb} = 0$ and $f = \tan h$ has the echo state property for all inputs if the following system converges to $z = 0$ uniformly for all input sequences $u^{+\infty}$ and state sequences $z^{+\infty}$:*

$$z_{k+1} = [(1 - a\Delta t)I + \Delta t L_k W]z_k, \tag{8}$$

where $\forall k \geq 0$, $z_k = x_k - y_k$ *with* $x^{+\infty}, y^{+\infty} \in X^{+\infty}$ *compatible with* $u^{+\infty}$ *and* $L_k = L_k(z_k, u_{k+1})$ *are diagonal matrices with entries in the interval* $(0, 1]$.

Proof. Note that when $W^{fb} = 0$, Eq. (7) becomes $x_{k+1} = (1 - a\Delta t)x_k + \Delta t f(W^{in}u^{\delta}_{k+1} + Wx_k)$. For any right infinite input sequence $(u^{\delta}_{k+1})_{k\geq 0} = u^{+\infty} \in U^{+\infty}$ and any two right infinite state vector sequences $x^{+\infty}, y^{+\infty} \in X^{+\infty}$ compatible with $u^{+\infty}$, we have:

$$\begin{aligned} x_{k+1} - y_{k+1} &= (1 - a\Delta t)x_k + \Delta t f(W^{in}u^{\delta}_{k+1} + Wx_k) \\ &\quad - (1 - a\Delta t)y_k - \Delta t f(W^{in}u^{\delta}_{k+1} + Wy_k). \end{aligned}$$

Applying the Mean-Value Theorem component-wise:

$$\begin{aligned} &f(Wx_k + W^{in}u^{\delta}_{k+1}) - f(Wy_k + W^{in}u^{\delta}_{k+1}) \\ &\quad = L_k(Wx_k + W^{in}u^{\delta}_{k+1} - Wy_k - W^{in}u^{\delta}_{k+1}) \\ &\quad = L_k W(x_k - y_k), \end{aligned}$$

where each $L_k = L_k(x_k, y_k, u_{k+1})$ is given by $L_k = diag(\ell^1_k, \ldots, \ell^N_k)$ with $0 < \ell^i_k \leq 1$ since the derivative of $\tan h$ is bounded between 0 and 1. Then we get,

$$x_{k+1} - y_{k+1} = (1 - a\Delta t)(x_k - y_k) + \Delta t L_k W(x_k - y_k).$$

Now, letting $z_k = x_k - y_k$, we obtain the system $z_{k+1} = [(1 - a\Delta t)I + \Delta t L_k W]z_k$. If one can show the uniform convergence of this system to $z = 0$ for all $u^{+\infty}$ and $z^{+\infty}$ then, by Theorem 2.1, the echo state property of the system given by Eq. (7) is satisfied for all inputs. □

The system given in Eq. (8) was studied in Hu and Wang (2002) (see Eq. 11 in Hu and Wang (2002)). Several sufficient conditions for the global asymptotic stability and global exponential stability (which means states approach the unique equilibrium point exponentially fast) were provided. Since global exponential stability implies uniform convergence, all the conditions provided in Hu and Wang (2002) for global exponential stability also imply the echo state property for the leaky integrator ESNs in Eq. (7) with no feedback. We provide two of those conditions (Theorem 6 and Corollary 10 in Hu and Wang (2002)) because of their similarity to the effective spectral radius.

Corollary 4.3. *If the spectral radius of the matrix* $\hat{M} \in \mathbb{R}^{N\times N}$ *defined below is smaller than* 1, *i.e.* $\rho(\hat{M}) < 1$, *then the echo state property for all inputs is satisfied for the leaky integrator ESN given in Eq. (7) with* $W^{fb} = 0$ *and* $f = \tan h$. $\hat{M}$ *given in* Box I *where* $|1 - a\Delta t + \hat{\ell}_i \Delta t w_{ii}| = \max_{0\leq \ell_i \leq 1} |1 - a\Delta t + \ell_i \Delta t w_{ii}|$.

A more restrictive condition can be given as:

Corollary 4.4. *If the spectral radius of the matrix* $\tilde{M} = \Delta t|W| + (1 - a\Delta t)I$ *where* $|W| = (|w_{ij}|)$, *is smaller than* 1, *i.e.* $\rho(\tilde{M}) < 1$, *then the leaky integrator ESN given in Eq. (7) with* $W^{fb} = 0$ *and* $f = \tan h$ *has the echo state property for all inputs.*

Note that the matrix $\tilde{M} = \Delta t|W| + (1 - a\Delta t)I$ differs from the matrix $\tilde{W} = \Delta t W + (1 - a\Delta t)I$ only by the involvement of the absolute value. Even though the effective spectral radius condition, i.e. $\rho(\tilde{W}) < 1$, is not sufficient for the echo state property, the new condition, i.e. $\rho(\tilde{M}) < 1$, is sufficient. This is similar to the standard ESN case where $\rho(W) < 1$ is not sufficient for the echo state property (as described in Section 3) but $\rho(|W|) < 1$ is sufficient as discussed in Section 4.1.

A simple recipe for the echo state property of the leaky integrator ESNs

The condition $\rho(\tilde{M}) < 1$ gives a simple way to construct internal weight matrices for the leaky integrator ESNs that satisfy the echo state property:

(i) Start with a random W with all non-negative entries, $w_{ij} \geq 0$ (note that $W = |W|$ in this case).
(ii) Scale W so that $\rho(\tilde{M}) < 1$ where $\tilde{M} = \Delta t W + (1 - a\Delta t)I$.
(iii) Change the signs of a desired number of entries of W to get negative connection weights as well.

Finally, we would like to point out that the new sufficient conditions are less restrictive than the existing $|1 - \Delta t(a - \sigma_{max})| < 1$ condition.

Example 4.2. Let us take $a = 1$, $c = 1$, $\delta = 0.5$ ($\Delta t = 0.5$) and $W = \begin{pmatrix} -0.9 & -1 \\ 0 & -0.9 \end{pmatrix}$ for the leaky integrator ESN given in Eq. (7) with $W^{fb} = 0$. The maximum singular value of W, $\sigma_{max} \approx 1.52$, gives $|1 - \Delta t(a - \sigma_{max})| \approx |1 - 0.5(1 - 1.52)| = 1.26 > 1$. Therefore, based on this condition, we cannot be sure that the system has the echo state property. However, using Eq. (9), $\hat{M} = \begin{pmatrix} 0.5 & 0.5 \\ 0 & 0.5 \end{pmatrix}$ where $\rho(\hat{M}) = 0.5 < 1$ actually proves the echo state property. This can also be seen from the more restrictive condition $\rho(\tilde{M}) = 0.95 < 1$ where $\tilde{M} = 0.5|W| + 0.5I = \begin{pmatrix} 0.95 & 0.5 \\ 0 & 0.95 \end{pmatrix}$.

5. A new definition for the echo state property

So far, we have investigated how the ESP can be lost for a spectral radius below unity, *in the case of zero input*. Furthermore, for zero input a spectral radius not exceeding unity is a necessary condition for the ESP. Obviously in practical applications one will usually have nonzero input. For nonzero input, the ESP can be obtained even with a spectral radius exceeding unity. A very simple example is the one-dimensional "reservoir" $x_{k+1} = \tan h(2x_k + 1)$ driven by constant input $u_k \equiv 1$. A quick look at a function plot (not shown) reveals that it has a global point attractor at $x \approx 0.995$ with uniform attraction from within $X = [-1, 1]$, thus has the ESP, but with a spectral radius of 2.

An intuitive explanation for the presence of ESP even at spectral radii beyond unity is that the input drives the reservoir units (which we assume here to be $\tan h$ units for the sake of discussion) toward the positive or negative branches of the sigmoid where stabilizing saturation effects start to become effective. In this context, it becomes apparent that the classical definition of the ESP, as in Definition 2.1, is not satisfactory. Consider a case where the input is some scalar random process which takes values in a compact interval, say u_k is uniformly distributed in $U = [-2, 2]$. A typical realization of the input process will have an expected absolute value of $E[|u_k|] = 1$, which would lead to the desired uniform convergence of state sequences in the sense of Theorem 2.1 for almost all input sequences even for a range of spectral radii beyond unity. However, for the input sequence $u_k \equiv 0$, which is a valid realization of the input process, the ESP would be lost for any spectral radius beyond 1. The original definition of the ESP takes account only of just the range U of admissible inputs, not of the distribution of the input process. In practice, it is however this distribution which determines the admissible range of spectral radius for *almost all* input sequences — which are the practically relevant ones, not the "pathological" ones, which destroy the ESP but occur with zero probability. Therefore, a more useful definition of the ESP is the following.

Definition 5.1 (*ESP Relative to an Input Process*)**.** Let $(U_k)_{k\in\mathbb{Z}}$ be a stochastic process, where the random variables U_k take values in a set U. A network $F : X \times U \to X$ satisfies the echo state property with respect to the process (U_k) if with probability one, for any left infinite input realization $u^{-\infty} \in U^{-\infty}$ and any two state vector sequences $x^{-\infty}, y^{-\infty} \in X^{-\infty}$ compatible with $u^{-\infty}$, it holds that $x_0 = y_0$.

However, a mathematical analysis of the implications of this definition needs a combination of tools from ergodic theory and non-autonomous dynamical systems. This is much harder than

$$\hat{M} = \begin{pmatrix} |1 - a\Delta t + \hat{\ell}_1 \Delta t w_{11}| & \Delta t |w_{12}| & \cdots & \Delta t |w_{1N}| \\ \Delta t |w_{21}| & |1 - a\Delta t + \hat{\ell}_2 \Delta t w_{22}| & \cdots & \Delta t |w_{2N}| \\ \vdots & \vdots & \ddots & \vdots \\ \Delta t |w_{N1}| & \Delta t |w_{N2}| & \cdots & |1 - a\Delta t + \hat{\ell}_N \Delta t w_{NN}| \end{pmatrix}, \tag{9}$$

Box I.

for the original definition of the ESP. Research in this direction is currently being pursued in the group of the second author.

6. Discussion

In this article we discussed, from various angles, the echo state property (ESP) and the closely related issue of the spectral radius of the reservoir weight matrix. The main technical contribution is a detailed analysis how the ESP is lost for specific weight patterns even when the spectral radius is below unity. Furthermore, we provided a novel algebraic criterion which is sufficient for the ESP for any input in reservoirs whose nonlinearity has a derivative bounded in $[-1, 1]$ (such as $\tanh$ reservoirs). We hope that the bifurcation analyses and link to the mathematically well-known concept of Schur stability will be interesting to both reservoir computing researchers and users.

Another motivation to write this paper was to clarify a number of problematic preconceptions which are rather widespread among researchers and engineers who use reservoir computing for their applications. The authors witness that a significant fraction of end-users of ESNs only consider reservoirs with a spectral radius below unity, following the (widespread but misguided) assumption that a spectral radius beyond unity destroys the ESP. One purpose of this article is to dissolve this misperception and encourage users to explore spectral radii greater than unity.

Conversely, the authors also witness examples of what one might call the inverse fallacy: by experimentation the spectral radius is scaled up to the point where the ESP is lost and a spectral radius slightly below this critical value is used. This leads to choices of spectral radii which are typically much greater than 1. The justification for this scheme is that in some research it has been found that reservoirs scaled to the "edge of chaos" give the best performance. This line of thought apparently originates – within the reservoir computing arena – in Bertschinger, Natschläger, and Legenstein (2004), Bertschinger and Natschläger (2004), where it was shown that, for reservoirs with binary threshold units, performance peaked when the reservoir weights were scaled to just below a critical value after which state convergence was lost. The authors of these papers were careful in stating the specific conditions under which these results were obtained and these results should not lead one to assume that scaling a reservoir to the borders of chaos is always beneficial.

In fact, there are tasks like the multistable switching circuits described in Jaeger (2002) (subsequently employed to model stable working memory mechanisms (Pascanu & Jaeger, 2011)) which require fast reaction times of reservoirs and fast locking into attractor states. Such behavior is best achieved with spectral radii much smaller than unity. Furthermore, it was shown later (Büsing, Schrauwen, & Legenstein, 2010) that binary reservoirs show a much more marked dependency of task performance on parameter scaling than the analog ESNs which are mostly used in applications.

In another study (Ozturk, Xu, & Principe, 2006), which explored the task performance of analog reservoirs with respect to specific information-theoretic metrics of the reservoir, it was found that performance is coupled to the average reservoir state entropy, which – on the one hand – increases when reservoirs are scaled toward the edge of chaos, but – on the other hand – can also be maximized by algebraically configuring the reservoir weight matrix such that its eigenvalues are spread as uniformly as possible about a circle in the complex plane. It was found here that task-optimal spectral radii varied, and were found to lie below 1 for all tasks considered, i.e. far away from chaos.

Finally, we would like to express our scepticism about the use of the term "edge of chaos" in the context of reservoir computing. This term was introduced in the seminal paper by Langton (1990) in the context of the emergence of universal computation in cellular automata. This is a setting which is at best indirectly related to reservoir computing, both with respect to the computational substrate (eminently discrete in cellular automata, typically analog in reservoirs) and with respect to the notion of "computation" (Turing computability in cellular automata vs. online signal processing in reservoirs). In fact, when the spectral radius of a reservoir weight matrix is scaled toward the critical value where the ESP is lost, one obtains a phase transition which typically does not lead into chaos but into some oscillatory mode. In the reservoir computing context it therefore is more appropriate to speak of an "edge of stability", as has been discussed in Verstraeten (2009).

To conclude, we briefly summarize a number of dispersed findings which amount to the following characterization of the current state of insight about the ESP in reservoir computing:

- The ESP is a property that is defined with respect to the reservoir *and* the nature of the driving input.
- The current definition of ESP is unsatisfactory in that it is uninformed about statistical properties of the input, which are however crucial in applications. An improved definition (Definition 5.1) is suggested, which however is mathematically difficult to analyze.
- The ESP (according to the new Definition 5.1) may be obtainable almost surely even for spectral radii (much) larger than unity if the driving input is sufficiently strong.
- While ensuring the ESP is mandatory for training and employing reservoirs, there is typically a wide range of spectral radii beyond 1 under which the ESP is obtained.
- There are no generally applicable recipes for the optimal setting of the spectral radius, in particular:
  - It is not required to scale the spectral radius below 1.
  - There is no general benefit in scaling the spectral radius toward the "edge of chaos".
- An appropriate setting of the spectral radius still has to be found by task-specific experimentation.

Appendix. Proof of Theorem 4.1

Proof. A straightforward proof can be given by noting that the diagonally Schur stable condition is shown in the text to be equivalent to the condition given in Buehner and Young (2006) for the echo state property (for the proof of this equivalence, see the Remark at the end of Section 4.1).

Another direct proof can be given using the Lyapunov theory.

Let us denote a finite input sequence $(u_1, \ldots, u_k) \in U^k$, $k \geq 1$ by $\bar{u}_k$ and the iterations of a point $x_0 \in X \subset \mathbb{R}^N$ under the echo state network $F : X \times U^k \to X$ by $x_k = F(x_0, \bar{u}_k)$. For the echo state property to hold, we want to show, by Theorem 2.1, that for

all $u^{+\infty} \in U^{+\infty}$, for all $x^{+\infty}$, $y^{+\infty} \in X^{+\infty}$ compatible with $u^{+\infty}$, the difference $\|x_k - y_k\|$ goes to zero uniformly. Note that since $X \times U^k$ is compact and $F : X \times U^k \to X$ is continuous, the image sets $X_k := \{F(x_0, \bar{u}_k) \mid x_0 \in X, \bar{u}_k \in U^k\}$ are compact. It also holds that $\forall k \geq 1, X_{k+1} \subseteq X_k \subseteq X_0 = X$ since any element $F(x_0, \bar{u}_{k+1}) \in X_{k+1}$ can be written as $F(x_0, \bar{u}_{k+1}) = F(F(x_0, u_1), (u_2, \ldots, u_{k+1})) \in X_k$.

Similarly, we define the difference sets $Z_k := \{x - y \mid x, y \in X_k\}$. Note that Z_k are compact and similarly we have $Z_{k+1} \subseteq Z_k \subseteq Z_0 = Z$.

We first define an equivalent system using the variable $z_k = x_k - y_k$ where $z_k \in Z_k$. Note that $z_{k+1} = x_{k+1} - y_{k+1} = f(Wx_k + W^{in}u_{k+1}) - f(Wy_k + W^{in}u_{k+1})$.

Using the Mean-Value Theorem component-wise:

$$\begin{aligned} &f(Wx_k + W^{in}u_{k+1}) - f(Wy_k + W^{in}u_{k+1}) \\ &\quad = L_k(Wx_k + W^{in}u_{k+1} - Wy_k - W^{in}u_{k+1}) \\ &\quad = L_k W(x_k - y_k), \end{aligned}$$

where each L_k is given by $L_k = diag(\ell_k^1, \ldots, \ell_k^N)$ with $0 < \ell_k^i \leq 1$ since the derivative of $\tan h$ is bounded between 0 and 1. Therefore, we can define the new system as:

$$z_{k+1} = L_k W z_k.$$

Note that $z = 0$ is an equilibrium point of this system and $L_k = L_k(x_k, y_k, u_{k+1})$. We need to show that all solutions converge to $z = 0$ *uniformly* for all (compatible) $z^{+\infty} \in Z^{+\infty}$ and $u^{+\infty} \in U^{+\infty}$.

Let us define the quadratic function $V : Z \to \mathbb{R}$, $V(z) = z^T P z$ where $P > 0$ is the diagonal positive matrix mentioned in the statement of the theorem. Note that $V(z) = 0 \Leftrightarrow z = 0$ and $V(z) > 0, \forall z \in Z \setminus \{0\}$, $V(z) \to \infty$ as $\|z\| \to \infty$ and

$$\begin{aligned} V(z_{k+1}) - V(z_k) &= (L_k W z_k)^T P (L_k W z_k) - z_k^T P z_k \\ &= z_k^T (L_k W)^T P (L_k W) z_k - z_k^T P z_k \\ &= z_k^T [(L_k W)^T P (L_k W) - P] z_k. \end{aligned}$$

To prove uniform global asymptotic stability (see for e.g. Khalil (2002)), one needs to find a positive definite, time independent function $N(z)$ such that $V(z_{k+1}) - V(z_k) = \Delta V(z) = z^T[(L_k W)^T P (L_k W) - P]z \leq -N(z)$ for all L_k and for all $z \in B_a(0)$ ($B_a(0)$ is a ball around zero).

By a simple calculation, one can see that $z_k^T[(L_k W)^T P (L_k W) - P]z_k \leq z_k^T[W^T P W - P]z_k$ for all z_k, i.e. the maximum is attained when L_k is the identity. Since $W^T P W - P$ is negative definite (by the assumption of the theorem), choosing $-N = W^T P W - P$, we prove the uniform global asymptotic stability of the origin. □

References

Bertschinger, N., & Natschläger, T. (2004). Real-time computation at the edge of chaos in recurrent neural networks. *Neural Computation*, *16*(7), 1413–1436.

Bertschinger, Nils, Natschläger, Thomas, & Legenstein, Robert A. (2004). At the edge of chaos: real-time computations and self-organized criticality in recurrent neural networks. In *Advances in neural information processing systems, Vol. 17 [Neural information processing systems, NIPS 2004, December 13–18, 2004]*. British Columbia, Canada: Vancouver.

Bhaya, Amit, & Kaszkurewicz, Eugenius (1993). On discrete-time diagonal and *d*-stability. *Linear Algebra and its Applications*, *187*, 87–104.

Büsing, L., Schrauwen, B., & Legenstein, R. (2010). Connectivity, dynamics, and memory in reservoir computing with binary and analog neurons. *Neural Computation*, *22*(5), 1272–1311.

Buehner, M., & Young, P. (2006). A tighter bound for the echo state property. *IEEE Transactions on Neural Networks*, *17*(3), 820–824.

Dominey, P. F. (1995). Complex sensory-motor sequence learning based on recurrent state representation and reinforcement learning. *Biological Cybernetics*, *73*, 265–274.

Hu, S., & Wang, J. (2002). Global stability of a class of discrete-time recurrent neural networks. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, *49*(8), 1104–1117.

Jaeger, H. (2001). The "echo state" approach to analysing and training recurrent neural networks. GMD report 148, GMD — german national research institute for computer science.

Jaeger, H. (2007). Echo state network. In *Scholarpedia*: *Vol. 2* (p. 2330).

Jaeger, H. (2010). Erratum note for the techreport: the "echo state" approach to analysing and training recurrent neural networks. Technical report.

Jaeger, H. (2002). Short term memory in echo state networks. *GMD-report 152, GMD — german national research institute for computer science*.

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

Jaeger, H., Lukosevicius, M., Popovici, D., & Siewert, U. (2007). Optimization and applications of echo state networks with leaky-integrator neurons. *Neural Networks*, *20*(3), 335–352. Echo State Networks and Liquid State Machines.

Kaszkurewicz, E., & Bhaya, A. (2000). *Matrix diagonal stability in systems and computation*. Birkhauser.

Khalil, H. K. (2002). *Nonlinear systems*. Prentice Hall.

Kuznetsov, I. U. A. (2004). Elements of applied bifurcation theory. In *Applied mathematical sciences*. Springer-Verlag.

Langton, C. G. (1990). Computation at the edge of chaos: phase transitions and emergent computation. *Physica D*, *42*(1–3), 12–37.

Lukoševičius, Mantas, & Jaeger, Herbert (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. http://www.cis.tugraz.at/igi/maass/psfiles/LSM-v106.pdf.

Ozturk, M. C., Xu, D., & Principe, J. C. (2006). Analysis and design of echo state networks for function approximation. *Neural Computation*, *19*, 111–138.

Packard, A., & Doyle, J. (1993). The complex structured singular value. *Automatica*, *29*(1), 71–109.

Pascanu, R., & Jaeger, H. (2011). A neurodynamical model for working memory. *Neural Networks*, *24*(2), 199–207. http://dx.doi.org/10.1016/j.neunet.2010.10.003.

Thompson, J. M. T., & Stewart, H. B. (2002). *Nonlinear dynamics and chaos*. Wiley.

Verstraeten, D. (2009). Reservoir computing: computation with dynamical systems. Ph.D. thesis, electronics and information systems. University of Ghent.