

A local Echo State Property through the largest Lyapunov exponent

Gilles Wainrib

Ecole Normale Supérieure, Département d'Informatique, Paris, France.

Mathieu N. Galtier

NeuroMathComp, Inria Sophia; UNIC, CNRS Gif; Minds, Jacobs University Bremen.

Abstract

Echo State Networks are efficient time-series predictors, which highly depend on the value of the spectral radius of the reservoir connectivity matrix. Based on recent results on the mean field theory of driven random recurrent neural networks, enabling the computation of the largest Lyapunov exponent of an ESN, we develop a cheap algorithm to establish a local and operational version of the Echo State Property.

Keywords: Reservoir computing, mean field theory, Lyapunov exponents, Echo State Networks.

1. Introduction

Echo State Networks (ESN) are neural networks designed for performing complex non-linear regression or classification tasks, such as non-linear time-series forecasting [1, 2]. As an instance of a more general framework called reservoir computing [3], the ESN architecture is based on a randomly connected recurrent neural network, called reservoir, which is driven by a temporal input. The state of the reservoir is a rich representation of the history of the inputs [4], so that a simple linear combination of the reservoir neurons is often a good predictor of the future of the inputs. The computation of the output connections can be done explicitly and corresponds to the minimization of the relative entropy between the network and the inputs dynamics [5], for which the associated gradient descent may be implemented with biologically plausible learning rules [6].

In this paper, we focus on the input-driven reservoir, which may be governed by a variety of dynamical systems beyond random neural networks [7], provided they produce consistent reservoir dynamics for a given input. This condition is of primary importance since its violation systematically leads to irrelevant results. In the original paper [1], Jaeger has given a condition, which he names *Echo State Property* (ESP), guaranteeing that the network states are consistent. This definition of the ESP and the equivalent formulations manipulate left infinite input time-series assuming that the initial condition occurs at $t = -\infty$. If n is the number of neurons in the reservoir, $\mathbf{x}(t) \in \mathbb{R}^n$ is the state of the reservoir at time $t \in \mathbb{Z}$ and $u(t) \in \mathbb{R}$ is the input to the reservoir of time $t \in \mathbb{Z}$. The ESP definition can be summarized as

Definition 1.1 (ESP [1]). *A network has the ESP if the network state $\mathbf{x}(t)$ is uniquely determined by any left-infinite input sequence $\{u(t-s) : s \in \mathbb{N}\}$.*

In other words, it means that the initial condition of the network (at $t = -\infty$) does not influence the trajectory of the states, which corresponds to the property that the input-driven network has a unique global attractor [8]. The ESP seems to be important in practice to design efficient reservoirs. Indeed, a network without ESP would have a poor accuracy in the inevitable presence of perturbations or noise: a small perturbation could bring the network to states it has never seen before, destroying the prediction capabilities of the network. Put differently, the network has to have some fading memory so that the initial conditions and perturbations do not impact the accuracy in the long term.

A fundamental result is that a bound on the maximum singular value η of the network connectivity matrix $\mathbf{J} \in \mathbb{R}^{n \times n}$ can provide the global ESP for every input. More specifically, if the dynamics of the network is governed by

$$\mathbf{x}_i(t+1) = S \left(\sum_{j=1}^n \mathbf{J}_{ij} \mathbf{x}_j(t) + \mathbf{m}_i u(t) \right) := G_i(\mathbf{x}(t), t) \quad (1)$$

where $\mathbf{m} \in \mathbb{R}^n$ is the input matrix, and $S(\cdot)$ is a sigmoid function with unit slope at the origin, then the following result holds:

Theorem 1.1 ([1]). *If $\eta < 1$, then the global ESP holds for every input.*

It is important to observe that the sufficient condition in 1.1 holds for the largest singular value η and not for the largest eigenvalue modulus ρ

(also called spectral radius), which are different for most matrices. Indeed, as pointed out in [9], the theory of random matrices gives a relationship between the maximum singular value η and the maximum eigenvalue ρ of the random matrix \mathbf{J} when the number of neurons tends to infinity. First, using recent results on the empirical spectral distribution of random matrices [10], one can show that large random matrices, whose entries are i.i.d. random variables with mean 0, finite variance $\frac{\sigma^2}{\sqrt{n}}$, have eigenvalues which tend to cover uniformly the disk of radius σ as the number of neurons tends to infinity. For these matrices, the non-scaled standard deviation of the weights σ is in fact equal to the spectral radius ρ . Second, one can use results concerning the right edge of the Marchenko-Pastur convergence [11, 12, 13] to show that $\eta \rightarrow 2\sigma$ when the number of neurons tends to infinity. From this result, as mentioned in [9], it is clear that the condition on the singular values translates to

Theorem 1.2. *When the number of neurons tends to infinity (and with the appropriate scaling of the weights variance by $\frac{1}{\sqrt{n}}$) the ESP holds for all inputs if $\rho = \sigma < 1/2$.*

Interestingly, there is here a clear gap between the theoretical sufficient condition $\eta < 1$ (i.e $\sigma < 1/2$) and the condition $\rho < 1$ (i.e $\sigma < 1$) which seems to be valid in practice [14]. Based on the notion of structured singular value and on concepts from control theory [15], a tighter sufficient condition has been derived involving the computation of the infimum of the maximal singular values of the connectivity matrix for variety of underlying norms [16]. Despite its improvement over the classical singular value, this criterion is difficult to compute in practice, remains poorly understood from the point of view of random matrix theory, and does not respond to the problem of finding a criterion which depends on input, as we will discuss below. It is also interesting to mention the recent work [9], where the concentration of measure phenomenon [17] is used to prove that:

Theorem 1.3 ([9]). *If $\rho < 1 - \epsilon$, then for any $\mathbf{x}, \tilde{\mathbf{x}} \in \mathbb{R}^n$, the probability that $\|G(\mathbf{x}, t) - G(\tilde{\mathbf{x}}, t)\| > \|\mathbf{x} - \tilde{\mathbf{x}}\|$ is exponentially small when the number of neurons is large.*

This result may seem sufficient to prove the contraction property with high probability, implying the ESP when $\sigma < 1$ with high probability. Actually, one must be careful because this result does not imply that $\mathbb{P}[\forall \mathbf{x}, \tilde{\mathbf{x}} \in \mathbb{R}^n, \forall t >$

0, $\|G(\mathbf{x}, t) - G(\tilde{\mathbf{x}}, t)\| > \|\mathbf{x} - \tilde{\mathbf{x}}\|$ is small with high probability, which is a much stronger result. However, the authors claim that their result shows why choosing σ close but smaller than one is sufficient in practice. In a sense, they argue that networks which do not verify criterion of Theorem 1.1 can still perform well in applications.

On the other side, it is also instructive to look for a necessary condition for the ESP. When the spectral radius ρ is larger than one, then the trivial null equilibrium of the system with zero input is linearly unstable, and Jaeger has shown that:

Theorem 1.4 ([1]). *When $\rho > 1$, the ESP does not hold for the null input.*

This result is in fact related to the existence of chaotic attractors as shown in [18]. Therefore, there is no hope for an ESP for all inputs beyond $\rho = 1$. However, in practice [14], it may be important to increase ρ above 1 to improve the ESN performance (to increase the memory for instance). If we want to go beyond $\rho = 1$, we need to drop the requirement to have the ESP for all inputs. It has recently been argued that one can define an ESP with respect to a particular input (or a set of inputs) [19]. Intuitively, this means that a network driven by an input will not display excessive irregularity if it has the ESP with respect to that input. In [19], a bound for the ESP is also provided

Theorem 1.5 ([19]). *If $\limsup_{j \rightarrow \infty} \sum_{i=-1}^{-j} (C_i - (1 + \ln(2))) I(C_i > 2) > \frac{\ln(\|\mathbf{J}\|)}{2}$, with C_i the smallest absolute component of the vector $\mathbf{mu}(i)$ and I is the indicator function, then the network has the ESP with respect to u .*

Intuitively, this bound plays with the saturation of the sigmoid and will be efficient if the inputs are strong enough to drive the network in the saturating regime. Although this is a loose bound, it has the interesting property that the network may have temporarily non-contracting dynamics and still have the ESP. These ideas are clearly related to the fact that stimulating a chaotic system can result in a synchronized non-chaotic response, as shown in the context of random neural networks in [20].

In this paper, we aim at contributing to the debate about the ESP using a mean-field approach applied to non-autonomous random neural networks in the large n limit. This theory derives a self-consistent statistical description of the reservoir dynamics unravelling the transition between regularity and irregularity in the network, based on a Lyapunov stability analysis. Although

brought very recently into the field of echo-state networks by [21], this theoretical approach has a long history, dating back to early works on spin-glass models [22, 23], followed by applications to random neural networks dynamics as in [18, 24, 25, 26]. The rigorous justification of this heuristic approach is non-trivial and has been resolved by [27, 28, 29] using large deviations techniques. These mathematical results actually requires to add an (arbitrary) small white-noise perturbation to the reservoir dynamics, in order to be able to use a change of probability formula (e.g. Girsanov Theorem) which is at the heart of the large deviation proof. The rigorous proof of the mean-field equations when this additional noise is removed remains open to our knowledge, but this is not a real problem in the ESN framework since adding such noise term is actually used in practice as a form of regularization, shown to be equivalent to the classical Tikhonov regularization [30].

The network we consider in this paper is a leaky integrator ESN [31] defined over a regular graph with degree αn , proportional to n . This means that every neuron in the network is only connected to αn other neurons, which is often used in practice to reduce computational complexity. To apply the mean-field theory, we will assume that n goes to infinity, but consider $\alpha \in (0, 1]$ to be a constant. The connections between neurons are weighted: we write \mathbf{J}_{ij} the weight from neuron j to neuron i . The weights are independent random variables satisfying:

$$\mathbb{E}(\mathbf{J}_{ij}) = 0 \quad \text{and} \quad \mathbb{E}(\mathbf{J}_{ij}^2) = \frac{\sigma^2}{n} < +\infty$$

This quenched hypothesis excludes any dynamics on the weights: they are kept constant after having been randomly drawn.

Given a one-dimensional input time series $u : \{1 \cdots T\} \rightarrow \mathbb{R}$, the classical neural network discrete dynamics is

$$\mathbf{x}_i(t+1) = (1 - l\tau)\mathbf{x}_i(t) + \tau S \left(\sum_{j \rightarrow i} \mathbf{J}_{ij} \mathbf{x}_j(t) + \mathbf{m}_i u(t) \right) \quad (2)$$

where $\mathbf{x}(t) \in \mathbb{R}^n$ corresponds to the activity of all the neurons in the network at time t . The vector of feedforward connections $\mathbf{m} \in \mathbb{R}^n$ is made of i.i.d. random variables satisfying $\mathbb{E}(\mathbf{m}_i) = 0$, $\mathbb{E}(\mathbf{m}_i^2) = m^2$. The numbers l and τ are in $[0, 1]$ and control the timescale of the ESN dynamics. The function $S(\cdot)$ is a typical odd sigmoid with $S(0) = 0$, $S'(0) = 1$, $S'(x) > 0$ and $xS''(x) \leq 0$. Note that it implies it is a 1-Lipschitz function. Actually, the

following computations become explicit when a particular choice is made: $S(x) = \text{erf}(\frac{\sqrt{\pi}}{2}x)$ (which follows the requirements above). We write $\sum_{j \rightarrow i}$ the summation of incoming information to a neuron which is only done over the neurons which are connected (through the graph) to the considered neuron.

The paper is organized as follows: in section 2, we derive a mean field theory of driven leaky integrator recurrent neural networks (RNNs) on a regular graph, and we show how it can be used to find the frontier between order and disorder for the network dynamics. Then, in section 3 we show how this can be used to define a computable condition guaranteeing an operational version of the ESP.

2. Mean-field theory for leaky ESN on regular graphs

2.1. Mean-field equations

From the seminal work [18], recently extended to the framework of stimulus driven RNN [20, 21], one can derive a self-consistent equation describing the statistical properties of the reservoir activity in the large n limit, which is known as the mean-field theory. In this section, we present an extension of [21] to leaky RNNs on regular graphs.

The key idea is to make the assumption that the variables $\mathbf{x}_i(t)$ are i.i.d. and independent of \mathbf{J} and \mathbf{m} . This makes possible to use the central limit theorem on $\sum_{j \rightarrow i} \mathbf{J}_{ij} \mathbf{x}_j(t)$ which can thus be considered as a Gaussian process. When $k = \alpha n \rightarrow +\infty$, all the $\mathbf{a}_i(t) = \sum_{j \rightarrow i} \mathbf{J}_{ij} \mathbf{x}_j(t) + \mathbf{m}_i u(t)$ for $i \in \{1..n\}$ tend to behave as centered Gaussian variables with variance

$$a^2(t) = \mathbb{E}[\mathbf{a}_i(t)^2] = \alpha \sigma^2 \gamma^2(t) + m^2 u(t)^2$$

where $\gamma^2(t)$ denotes the variance of $\mathbf{x}_i(t)$ (independent of i). The iteration equation $\mathbf{x}_i(t+1) = (1-l\tau)\mathbf{x}_i(t) + \tau S(\mathbf{a}_i(t))$ is going to help us derive the mean-field dynamical system describing the variance of the \mathbf{x}_i . However, the independence between $\mathbf{x}_i(t)$ and $S(\mathbf{a}_i(t))$ is not granted and we cannot simply add their variance. Nonetheless, we can compute

$$\gamma^2(t+1) = (1-l\tau)^2 \gamma^2(t) + \tau^2 F(a^2(t)) + 2\tau(1-l\tau)R(t,t) \quad (3)$$

with

$$F(z^2) = (2\pi)^{-1/2} \int_{\mathbb{R}} S^2(zx) e^{-x^2/2} dx = \frac{2}{\pi} \arcsin\left(\frac{\pi z^2}{2 + \pi z^2}\right) \quad (4)$$

according to the technical result in the appendix of [32], and

$$R(s, t) = \mathbb{E}[\mathbf{x}_i(s)S(\mathbf{a}_i(t))] = (1 - l\tau)R(s - 1, t) + \tau Q(s - 1, t) \quad (5)$$

where

$$Q(s, t) = \mathbb{E}[S(\mathbf{a}_i(s))S(\mathbf{a}_i(t))] \quad (6)$$

Using again the result in [32], we can show that

$$\begin{aligned} Q(s, t) &= G\left(C(s, t), \gamma^2(s), \gamma^2(t)\right) \\ &= \frac{2}{\pi} \sin^{-1} \left(\frac{\pi}{2} \frac{\alpha\sigma^2 C(s, t) + m^2 u(s)u(t)}{\sqrt{\left(1 + \frac{\pi}{2}a^2(s)\right)\left(1 + \frac{\pi}{2}a^2(t)\right)}} \right) \end{aligned} \quad (7)$$

where

$$C(s, t) = \mathbb{E}[\mathbf{x}_i(s)\mathbf{x}_i(t)] = (1 - l\tau)C(s, t - 1) + \tau R(s, t - 1) \quad (8)$$

The recursive combination of equations (3), (5) and (8) provides a consistent description of the global variance of the neurons. An algorithm is provided in algorithm 1.

2.2. Order-disorder transition

The consistency equation (3) characterizes the transition between order and disorder in the network as a function of the variance of the connections σ^2 and the sparsity coefficient α . We first illustrate this phenomenon in the autonomous case and then discuss its impact in the input driven case.

2.2.1. Without input

the terms $\mathbf{x}_i(t)$ and $S(\mathbf{a}_i(t))$ are independent, and the third term in (3) disappears. Thus, let us study the autonomous dynamical system $\gamma^2(t+1) = (1 - l\tau)^2 \gamma^2(t) + \tau^2 F(\alpha\sigma^2 \gamma^2(t))$. Due to the properties of the sigmoid function S , the function F is increasing, concave and satisfies $F(0) = 0$ and $F'(0) = 1$. Therefore, the function $\Psi : x \mapsto (1 - l\tau)^2 x + \tau^2 F(\alpha\sigma^2 x)$ is also increasing and concave. Therefore, the slope at 0, denoted $\mu = \Psi'(0)$, is the effective parameter controlling the phase transition, and is given by

$$\mu = (1 - l\tau)^2 + \tau^2 \alpha \sigma^2 \quad (9)$$

This leads to a simple characterization of the behavior of the system for different values of μ :

- $\gamma^2(t)$ converges to $\gamma_\infty^2 = 0$ if $\mu < 1$
- $\gamma^2(t)$ converges to a limit value $\gamma_\infty^2 > 0$ if $\mu > 1$

In the first situation $\sigma < \sigma^* = \sqrt{\frac{l}{\alpha}(\frac{2}{\tau} - l)}$, all neuron variables converge to the quiescent state, whereas the network behavior becomes irregular as soon as $\sigma > \sigma^*$. Note that this generalizes the classical results of [18, 24] dealing with the case $\tau = \alpha = l = 1$, a case which is also treated in [33], where stability criteria are established for dynamical systems defining recurrent kernels for infinite-dimensional ESN.

2.2.2. With inputs, largest Lyapunov exponent

When the system is driven by external inputs, the network will never go to a quiescent state. Indeed, it is clear from equation (3) that the situation $\gamma^2(t) = 0$ will never happen. But one should not conclude that the network is always disordered because it could be strongly locked to the inputs, which is another way of defining the notion of order in such systems. The network will be said to be in order (resp. disorder) when a small perturbation independent of the inputs will vanish (rep. grow) with time. This corresponds to the notion of Lyapunov stability for the input driven system. The largest Lyapunov is below 1 in the case of robustness of the dynamics to small perturbations (order), and above 1 when the dynamics is significantly impacted by small perturbations, as is the case in chaotic systems (disorder). Formally, the largest lyapunov exponent can be defined as:

$$\lambda[u] := \lim_{t \rightarrow \infty, \delta(0) \rightarrow 0} \left(\frac{\delta^2(t)}{\delta^2(0)} \right)^{1/t} \quad (10)$$

where $\delta(t)$ is a distance at time t between two trajectories of (2) starting with different initial conditions separated by $\delta(0)$. More precisely, let us define $\delta(t)$ such that $\mathbf{x}_i(t) - \mathbf{x}'_i(t) \sim \mathcal{N}$ where \mathbf{x}_i and \mathbf{x}'_i are two solutions of (2) starting from two different initial conditions with $\mathbf{x}_i(0) - \mathbf{x}'_i(0) \sim \mathcal{N}(0, \delta(0)^2)$. In the situation where $\delta(t)$ is small, we have the following recurrence equation:

$$\begin{aligned} & \mathbf{x}_i(t+1) - \mathbf{x}'_i(t+1) \\ = & (1 - l\tau)(\mathbf{x}_i(t) - \mathbf{x}'_i(t)) + \tau (S(\mathbf{a}_i(t)) - S(\mathbf{a}'_i(t))) \\ = & (1 - l\tau)(\mathbf{x}_i(t) - \mathbf{x}'_i(t)) \\ + & \tau S'(\mathbf{a}_i(t)) \left(\sum_{j \rightarrow i} \mathbf{J}_{ij}(\mathbf{x}_j(t) - \mathbf{x}'_j(t)) \right) + o(\delta(t)) \end{aligned}$$

Therefore, one obtains the following relationship on the variances:

$$\begin{aligned}\delta^2(t+1) &= (1-l\tau)^2\delta^2(t) \\ &+ \tau^2\alpha\sigma^2\Phi(\alpha\sigma^2\gamma^2(t)+m^2u(t)^2)\delta^2(t) \\ &+ o(\delta^2(t))\end{aligned}\tag{11}$$

with

$$\Phi(z^2) := (2\pi)^{-1/2} \int S'^2(zx)e^{-x^2/2}dx = \frac{1}{\sqrt{1+\pi z^2}}.\tag{12}$$

When $\gamma^2(t)$ is obtained by solving iteratively (3), one can find the local Lyapunov exponent:

$$\lambda(t) := (1-l\tau)^2 + \tau^2\alpha\sigma^2\Phi(\alpha\sigma^2\gamma^2(t)+m^2u(t)^2)\tag{13}$$

When $\lambda(t)[u] < 1$, local asymptotic stability is ensured and the reservoir tends to be synchronized by the input, whereas when $\lambda(t)[u] > 1$, small perturbations are exponentially amplified and the reservoir is likely to enter a chaotic regime. It is natural that this measure depends on time because, for instance in the case $\sigma > 1$, synchronized states will only appear during periods when the input is sufficiently large compared to σ .

Combining (10) and (11), one can define a global finite horizon largest Lyapunov exponent as:

$$\lambda_T[u] := \left(\prod_{t=1}^T \lambda(t) \right)^{\frac{1}{T}}\tag{14}$$

where $\lambda(t)$ is defined in (13). Furthermore, at this stage, one already obtains an important property, showing that adding external input can only stabilize the system. Indeed, since $\Phi \leq 1$ (due to the fact that $|S'| \leq 1$), we always have the following inequality:

$$\lambda_T \leq \mu\tag{15}$$

Therefore, if the system without external input is in the ordered phase, namely when $\mu < 1$, then it is also in the ordered phase ($\lambda_T[u] < 1$) for all input. This results supports the fact that, in practice, $\rho < 1$ is a sufficient condition for the ESP.

In figure 1, we have applied algorithm 1 to estimate Λ_T in the case where $u(t) = 0$ (left) and where $u(t) = \sin(\omega t)$ (right) for various values of parameters σ and τ . In this figure, one observes that Λ_T is an increasing function of

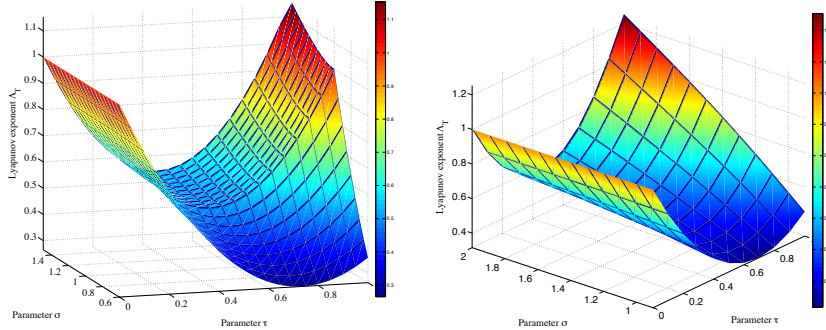


Figure 1: Numerical estimation of the global largest Lyapunov exponent Λ_T using algorithm 1 as a function of τ and σ . **Left:** Input $u(t) = 0$. **Right:** Input $u(t) = \sin(\omega t)$ with $\omega = 0.25$. Other parameters: $T = 1000$, $l = \alpha = m = 1$.

σ , which is a consequence of the fact that both $\gamma(t)$ and $\lambda(t)$ are increasing functions of σ , and corresponds with the intuition that increasing the disorder level would increase the instability of the dynamics. The case of null input (left) with $\tau = 1$ corresponds to the classical case [25], and displays a kink at $\sigma = 1$, whose consequences in terms of information processing has been discussed in [34]. However, the impact of the leak rate τ on the Lyapunov exponent has not been studied so far to our knowledge, and reveals an interesting U-shaped behavior indicating that there exists an optimal intermediate value of τ which minimizes the instability of the system. Our purpose in the present paper is to evaluate the Lyapunov exponent when the system is driven by an external time-series, which is displayed on the right panel of figure 1 with $u(t) = \sin(\omega t)$. This figure shows that the overall behavior is similar to the null-input case, with the expected difference that σ must be set much larger than one (around 1.6 when $\tau = 1$) to observe an exponent $\Lambda_T > 1$. Intuitively, the driven system is more stable because the input acts as a time-dependent bias in the sigmoid transfer function, hence reducing its average slope $|S'|$ along a trajectory, and therefore the norm of the Jacobian matrix which controls the local expansion rate. Notice that the quantity Φ defined in (12) corresponds to the average squared slope of S , where the average will be taken with respect to the Gaussian distribution with appropriate time-dependent variance (13).

3. Local Echo State Property

In this section, we discuss in more details the connection between the Lyapunov exponent Λ_T and the ESP.

3.1. Definition

The intuition behind the ESP is that the network should follow a reproducible and robust attractor. If the attractor is not stable, then the output connectivity matrix would be learned on a trajectory which could be different from the trajectory observed during the prediction or test phase, leading to poor accuracy. A key element to quantify the stability of the network trajectory is to measure the impact of small perturbations. If these perturbations are amplified over time then the dynamics is too irregular for good performance, the network is chaotic. Therefore, we define a local version of the ESP which guarantees the robustness of the dynamics to perturbations:

Definition 3.1 (Local ESP). *A driven dynamical system has the local Echo State Property if a small perturbation $\tilde{\mathbf{x}}(t_0) = \mathbf{x}(t_0) + \delta$ applied at time t_0 decreases to 0 in the large time asymptotic limit, namely $\|\mathbf{x}(t) - \tilde{\mathbf{x}}(t)\| \rightarrow 0$ when $t \rightarrow \infty$ for δ sufficiently small.*

This definition differs from the traditional ESP Definition 1.1 in two aspects: first, it deals with perturbations which do not necessarily occur at time $t = -\infty$. This definition only asks the perturbed solution to converge eventually towards the unperturbed solution, whereas the traditional definition asks that the solutions are identical based on the fact that the perturbation occurred an infinite number of time steps before. This definition is closer to the practical application of ESN where the initial condition corresponds to $t = 0$. Second, this definition only guarantees a local stability of the trajectories asking them to be robust only to small enough perturbations. On the other hand the traditional ESP requires that even large perturbations leave the trajectory unchanged. Put differently the traditional ESP guarantees a unique globally stable attractor, whereas the local ESP guarantees local stability of possibly many attractors (which have the same statistical properties).

We claim that the local ESP is sufficient for the good behavior of the network for most applications. More precisely, the only danger for systems that satisfy the local ESP, and not the traditional global ESP, is when learning is made on one attractor and prediction / test is made on another. In applications, if prediction / test is made immediately after learning such that

we are sure to stay on the same attractor, then the local ESP is sufficient. On the other hand, if the initialization of the prediction / test phase is done randomly, then the network may converge to a different attractor than that explored during learning. In that case, one would expect the performance to be poor.

3.2. Characterization

Measuring the evolution of small perturbations precisely corresponds to computing the largest Lyapunov exponent. Indeed, if $\lambda < 1$ then a small perturbation will eventually vanish and the perturbed solution will converge to the unperturbed solution. Therefore, by construction we have the following quantitative criterion for the local ESP:

Theorem 3.1. *If $\lambda[u] < 1$ then the network has the local ESP for the input u .*

Some remarks:

- The local ESP can be valid for systems experiencing temporary growth of perturbations as long as they are followed by a more important decrease. What matters in the definition of the local ESP is the balance of growth and decrease over a long time.
- From the key inequality (15), we deduce that the local ESP hold for all inputs whenever $\mu < 1$. This is a further argument supporting the practical criterion of a spectral radius below 1 should work for all inputs.
- There is a unique σ_L such that $\lambda(\sigma = \sigma_L) = 1$ and that the local ESP holds for all $\sigma < \sigma_L$. Indeed, we claim first that for any input u , the mapping $\sigma^2 \mapsto \lambda[u]$ is increasing. The proof is as follows. The function F is increasing, concave with $F'(0) = 1$. Therefore, equation (3) shows that $\gamma(t)$ increases sublinearly with σ^2 . Performing a simple change of variable in equation (12), it is easy to see that $\Phi(z^2)$ decreases slower than $1/z$ when z^2 increases. Therefore, $\sigma^2 \Phi(\alpha \sigma^2 \gamma^2(t) + m^2 u(t)^2)$ increases with σ^2 and so does λ_T according to equation (14). Finally, one observes that $\lambda(\sigma = 0) \leq 1$ and $\lambda(\sigma = +\infty) = +\infty$.

3.3. Numerical experiments

We now present an algorithm to compute $\lambda[u]$. A dichotomy algorithm, or any zero search algorithm for non-linear functions, could be implemented to find an approximation of σ_L , but given the cheap computational cost of computing $\lambda[u]$ for any σ , we will rather perform a grid search in this paper.

The algorithm to compute $\lambda_T[u]$ is stated below, when $\sigma, m, \alpha, l, \tau$ and u have been fixed. Note that this algorithm is computationally cheap, in $O(T)$,

Algorithm 1 Computing $\lambda_T[u]$

```

1:  $\lambda \leftarrow 1$ 
2:  $\gamma^2 \leftarrow 0$ 
3:  $R, C, \gamma_{\text{hist}} \leftarrow 0 \in \mathbb{R}^k$ 
4: for  $t = 1 : T$  do
5:   for  $s = 1 : k-1$  do
6:      $C[s] \leftarrow (1 - l\tau)C[s + 1] + \tau R[s + 1]$ 
7:      $R[s + 1] \leftarrow (1 - l\tau)R[s] + \tau G(C[s], \gamma^2, \gamma_{\text{hist}}[s])$ 
8:   end for
9:    $a \leftarrow \alpha\sigma^2\gamma^2 + m^2u(t)^2$ 
10:   $\lambda \leftarrow \lambda((1 - l\tau)^2 + \tau^2\alpha\sigma^2\Phi(a))^{1/T}$ 
11:   $\gamma^2 \leftarrow (1 - l\tau)^2\gamma^2 + \tau^2F(a) + 2\tau(1 - l\tau)R[-1]$ 
12:   $\gamma_{\text{hist}}[: -1] \leftarrow \gamma_{\text{hist}}[1 : ]$ 
13:   $\gamma_{\text{hist}}[-1], C[k] \leftarrow \gamma^2$ 
14: end for
15: return  $\lambda$ 

```

especially compared to the simulation of the full network.

To show on a numerical example that the local ESP guarantees good accuracy, we have computed the prediction performance for a prediction task. More precisely, we consider here the classical task of Mackey-Glass (MG) time-series prediction. The MG dynamical system [35] is given by the following delayed differential equation:

$$\dot{u}(t) = -bu(t) + \frac{au(t - \delta_{MG})}{1 + u(t - \delta_{MG})^{10}} \quad (16)$$

For each time-series, the task is to predict $u(t + 1)$ (one-step ahead) given the past $u(1), \dots, u(t - 1), u(t)$. Training is done on half of the time-series and predictions are made for the other half. For different variances of the

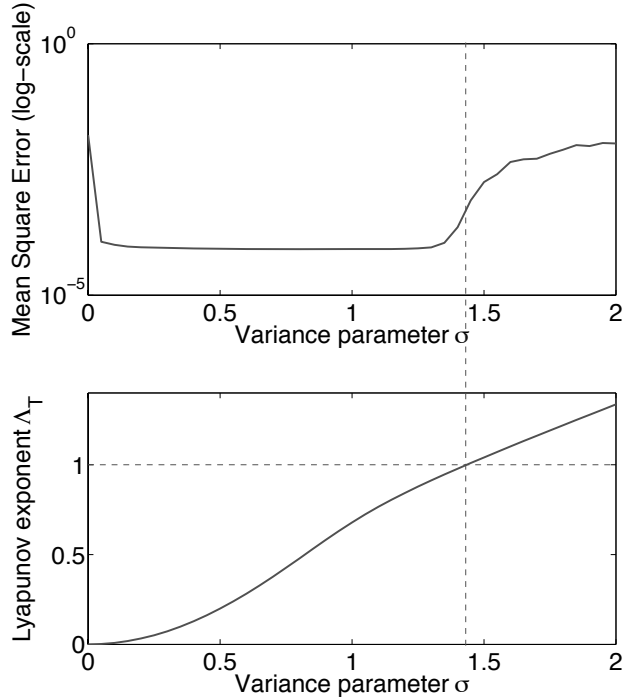


Figure 2: **Top:** this figure displays the prediction accuracy (mean-square error on a testing set) of an ESN when σ vary from 0 to 2. **Bottom:** this figure display the Lyapunov exponent Λ_T as a function of σ . The dashed lines help seeing the critical value $\sigma^* \simeq 1.57$ which both corresponds to $\Lambda_T = 1$ and to the transition to a regime of poor accuracy for the ESN. For the simulations the parameters were $n = 2000$, $\alpha = l = \tau = 1$, $m = 1$, and $T = 2000$. The time-series to predict is a solution of the Mackey-Glass chaotic dynamical system with $\delta_{MG} = 18$.

recurrent weights, we have plotted the accuracy of an ESN in figure 2 (top). This accuracy corresponds to the quantity $H = \frac{1}{T} \sum_{t=0}^{T-1} (u(t+1) - \mathbf{w}' \cdot \mathbf{x}(t))^2$, where $\mathbf{w} \in \mathbb{R}^n$ was computed with the usual Wiener-Hopf solution: $\mathbf{w} = \left(\sum_{t=0}^{T-1} \mathbf{x}(t) \cdot \mathbf{x}(t)' \right)^{-1} \cdot \left(\sum_{t=0}^{T-1} \mathbf{x}(t) u(t+1) \right)$. We see that even for some $\sigma > 1$ the accuracy is good although the global ESP for all inputs is not satisfied any more. However, the accuracy becomes significantly poorer after a certain critical value for σ . In figure 2 (bottom), we have plotted the value of the Lyapunov exponent Λ_T computed with the algorithm above. We see that it crosses 1 quite precisely at a critical value σ^* for which the accuracy moves to a regime of much higher values.

In order to further investigate the link between the local Lyapunov exponent Λ_T and ESN performance, we have generated several discrete time-series corresponding to various values of $\delta_{MG} \in \{10, 12, 14, 16, 18, 20, 22\}$ with parameters $a = 0.2$ and $b = 0.1$. In figure 3 (right), ESN performance is measured by the Mean Square Error on a testing set and is displayed as a function of the variance parameter σ , for various values of the delay δ_{MG} . Good performance is typically achieved for an intermediate range of values of σ , and one observes that the upper value of this range is smaller for higher values of δ_{MG} , as indicated by the black arrow. We interpret this loss of performance for high values of σ as related to the loss of the ESP. If this is indeed the case, then it should be possible to predict this behavior by using Algorithm 1 to compute σ^* , the value of σ for which local Lyapunov exponent Λ_T becomes larger than 1. As displayed in figure 3 (left), σ^* is a decreasing function of δ_{MG} , which is perfectly consistent with the above observation. This numerical example illustrates that the proposed theoretical advance presented in this article helps predicting and understanding the behavior of the performance curve for Echo-State Networks. However, finding the optimal value of all the hyper-parameters, beyond a systematic cross-validation procedure, remains a challenging theoretical problem.

4. Conclusion

In this paper, we have shown that the mean field theory for ESN developed in [21] can be, first, extended to leaky integrator networks on regular graphs; and, second, used to compute accurately a condition for the local ESP corresponding to the edge of chaos. We argue that the local ESP with respect to the given input is the useful condition to check in many applications, to ensure that the ESN representation is stable to small perturbations. We do not claim that the edge of chaos is always the best regime, but it has been shown that for some applications, typically requiring a lot of memory, it was optimal [36]. We believe that the proposed method to assess the local ESP should be systematically used to make sure the ESN has a regular dynamics leading to good accuracy. However, finding the optimal values of the hyper-parameters (e.g. σ, τ, α etc.) for a given supervised learning task necessitates to take into account both the input and the target, which goes beyond the scope of the present approach : we provide a method to compute a bound for these parameters, given the input time-series, to ensure the ESP.

The theory has only been detailed for one dimensional inputs, but the

extension of this approach to multidimensional inputs is not difficult (see [21]). Extending this method to other types of dynamics should be feasible as long as the computation of F and Φ can be numerically done or conveniently reformulated. Finally, the mean-field approach only deals with the limit of very large networks $n \rightarrow \infty$, whereas in practice the aim might be to perform a given task with the smallest possible reservoir to avoid over-fitting issues. Therefore, a further investigation of the finite-size effects around the mean-field limit would be of interest. For instance, a related question has been studied in [37], where it is shown that networks with a variance parameter $\sigma < 1$ have a probability to be unstable which is maximal for a specific size of the reservoir.

References

References

- [1] H. Jaeger, Short term memory in echo state networks, GMD-Forschungszentrum Informationstechnik, 2001.
- [2] H. Jaeger, H. Haas, Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication, *Science* 304 (5667) (2004) 78–80.
- [3] M. Lukosevicius, H. Jaeger, Survey: Reservoir computing approaches to recurrent neural network training, *Computer Science Review* 3 (3) (2009) 127–149.
- [4] D. V. Buonomano, M. M. Merzenich, Temporal information transformed into a spatial code by a neural network with realistic properties, *Science* (1995) 1028–1028.
- [5] M. Galtier, C. Marini, G. Wainrib, H. Jaeger, Relative entropy minimizing noisy non-linear neural network to approximate stochastic processes., Submitted. Preprint: arXiv:1402.1613.
- [6] M. Galtier, G. Wainrib, A biological gradient descent for prediction through a combination of stdp and homeostatic plasticity, *Neural computation* 25 (11) (2013) 2815–2832.
- [7] J. Dambre, D. Verstraeten, B. Schrauwen, S. Massar, Information processing capacity of dynamical systems, *Scientific reports* 2.

- [8] D. N. Cheban, Global attractors of non-autonomous dissipative dynamical systems, (Interdisciplinary mathematical sciences, Vol. 1), World Scientific, River Edge, NJ, 2004.
- [9] B. Zhang, D. J. Miller, Y. Wang, Nonlinear system modeling with random matrices: echo state networks revisited, *Neural Networks and Learning Systems*, IEEE Transactions on 23 (1) (2012) 175–182.
- [10] T. Tao, V. Vu, M. Krishnapur, Random matrices: Universality of esds and the circular law, *The Annals of Probability* 38 (5) (2010) 2023–2065.
- [11] V. A. Marcenko, L. A. Pastur, Distribution of eigenvalues for some sets of random matrices, *Sbornik: Mathematics* 1 (4) (1967) 457–483.
- [12] S. Geman, A limit theorem for the norm of random matrices, *The Annals of Probability* 8 (2) (1980) 252–261.
- [13] Z. Bai, J. W. Silverstein, *Spectral analysis of large dimensional random matrices*, Springer, 2010.
- [14] M. Lukosevicius, A practical guide to applying echo state networks, in: *Neural Networks: Tricks of the Trade*, Springer, 2012, pp. 659–686.
- [15] W. Lohmiller, J.-J. E. Slotine, On contraction analysis for non-linear systems, *Automatica* 34 (6) (1998) 683–696.
- [16] M. Buehner, P. Young, A tighter bound for the echo state property, *Neural Networks*, IEEE Transactions on 17 (3) (2006) 820–824.
- [17] M. Ledoux, *The concentration of measure phenomenon*, Vol. 89, AMS Bookstore, 2005.
- [18] H. Sompolinsky, A. Crisanti, H. Sommers, Chaos in random neural networks, *Physical Review Letters* 61 (3) (1988) 259–262.
- [19] G. Manjunath, H. Jaeger, Echo state property linked to an input: Exploring a fundamental characteristic of recurrent neural networks, *Neural computation* 25 (3) (2013) 671–696.
- [20] K. Rajan, L. Abbott, H. Sompolinsky, Stimulus-dependent suppression of chaos in recurrent neural networks, *Physical Review E* 82 (1) (2010) 011903.

- [21] M. Massar, S. Massar, Mean-field theory of echo state networks, *Physical Review E* 87 (4) (2013) 042809.
- [22] H. Sompolinsky, A. Zippelius, Dynamic theory of the spin-glass phase, *Physical Review Letters* 47 (5) (1981) 359.
- [23] H. Sompolinsky, A. Zippelius, Relaxational dynamics of the edwards-anderson model and the mean-field theory of spin-glasses, *Physical Review B* 25 (11) (1982) 6860.
- [24] B. Cessac, B. Doyon, M. Quoy, M. Samuelides, Mean-field equations, bifurcation map and route to chaos in discrete time neural networks, *Physica D: Nonlinear Phenomena* 74 (1) (1994) 24–44.
- [25] L. Molgedey, J. Schuchhardt, H. G. Schuster, Suppressing chaos in neural networks by noise, *Phys. Rev. Lett.* 69 (1992) 3717–3719. doi: 10.1103/PhysRevLett.69.3717.
URL <http://link.aps.org/doi/10.1103/PhysRevLett.69.3717>
- [26] O. Faugeras, J. Touboul, B. Cessac, A constructive mean-field analysis of multi-population neural networks with random synaptic weights and stochastic inputs, *Frontiers in computational neuroscience* 3.
- [27] G. B. Arous, A. Guionnet, Large deviations for langevin spin glass dynamics, *Probability Theory and Related Fields* 102 (4) (1995) 455–509.
- [28] O. Moynot, M. Samuelides, Large deviations and mean-field theory for asymmetric random recurrent neural networks, *Probability Theory and Related Fields* 123 (1) (2002) 41–75.
- [29] T. Cabana, J. Touboul, Large deviations, dynamics and phase transitions in large stochastic heterogeneous neural networks, arXiv preprint arXiv:1302.6951.
- [30] C. M. Bishop, Training with noise is equivalent to tikhonov regularization, *Neural computation* 7 (1) (1995) 108–116.
- [31] H. Jaeger, M. Lukosevicius, D. Popovici, U. Siewert, Optimization and applications of echo state networks with leaky-integrator neurons, *Neural Networks* 20 (3) (2007) 335–352.

- [32] C. K. Williams, Computation with infinite neural networks, *Neural Computation* 10 (5) (1998) 1203–1216.
- [33] M. Hermans, B. Schrauwen, Recurrent kernel machines: Computing with infinite echo state networks, *Neural Computation* 24 (1) (2012) 104–133.
- [34] T. Toyozumi, L. Abbott, Beyond the edge of chaos: Amplification and temporal integration by recurrent networks in the chaotic regime, *Physical Review E* 84 (5) (2011) 051908.
- [35] M. Mackey, L. Glass, *Science* 197 (287).
- [36] N. Bertschinger, T. Natschläger, Real-time computation at the edge of chaos in recurrent neural networks, *Neural computation* 16 (7) (2004) 1413–1436.
- [37] G. Wainrib, L. C. G. del Molino, Optimal system size for complex dynamics in random neural networks near criticality, *Chaos: An Interdisciplinary Journal of Nonlinear Science* 23 (4) (2013) –. doi:<http://dx.doi.org/10.1063/1.4841396>.

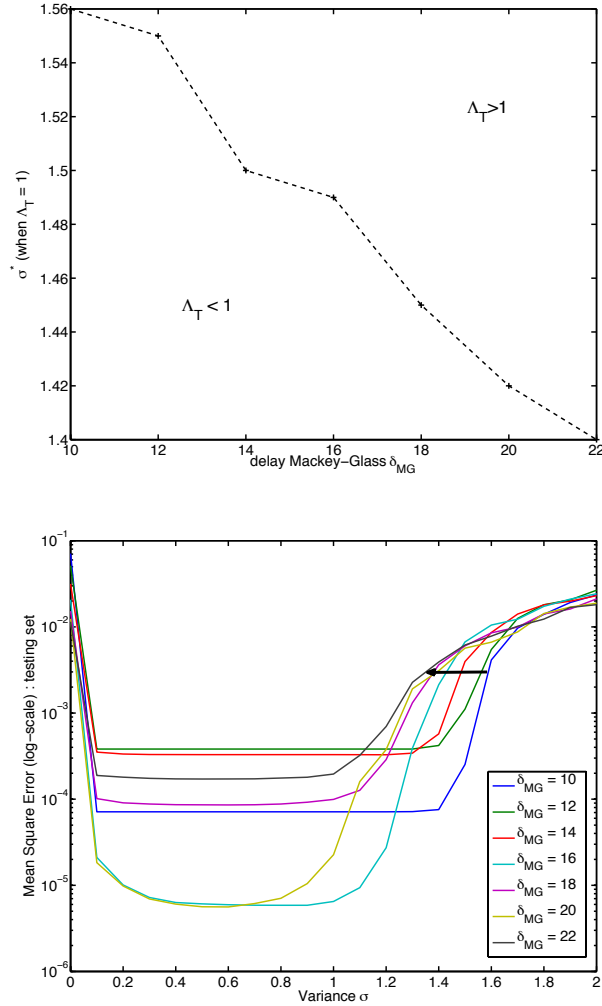


Figure 3: Analysis of ESN performance as a function of the delay δ_{MG} in the Mackey-Glass prediction task. **Left:** Using Algorithm 1, we were able to compute the value σ^* of the weights variance for which Λ_T becomes larger than 1, corresponding to the edge of chaos. The value of σ^* depends on the delay parameter δ_{MG} : it appears that increasing δ_{MG} leads to a smaller value of σ^* . **Right:** Mean-square error (testing set) as a function of the variance σ for the Mackey-Glass prediction task, for different values of the delay δ_{MG} . This figure confirms the prediction made in the Left panel : as indicated by the **black arrow**, for higher values of δ_{MG} , the value of σ where the performance starts to become poorer appears earlier. For the simulations the parameters of the MG system were $a = 0.2$, $b = 0.1$ with a time-step $\delta t = 1$ and the ESN parameters were $n = 100$, $\alpha = l = \tau = 1$, $m = 1$, for time-series of length $T = 2000$.