Neural Networks 76 (2016) 39–45

Contents lists available at ScienceDirect

Neural Networks

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

A local Echo State Property through the largest Lyapunov exponent

Gilles Wainrib [a,*], Mathieu N. Galtier [b]

[a] Ecole Normale Superieure, Departement d'Informatique, Paris, France
[b] UNIC, CNRS Gif sur Yvette, France

ARTICLE INFO

Article history:
Received 2 June 2015
Received in revised form 29 November 2015
Accepted 23 December 2015
Available online 13 January 2016

Keywords:
Reservoir computing
Mean field theory
Lyapunov exponents
Echo State Networks

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.

1. Introduction

Time series prediction is becoming ubiquitous in science and technology. Interestingly, common machine learning algorithms, such as feedforward neural networks (LeCun, Bengio, & Hinton, 2015), are not designed to naturally process random variables whose samples are not independently and identically distributed (Haykin, 2009). Times series are such random variables since each time step is highly correlated to the previous.

Recurrent neural networks are more naturally associated to times series since they both share the same nature: they are a trajectory of a (possibly stochastic) dynamical system (Funahashi & Nakamura, 1993). However, the learning process of the network recurrent weights can be difficult in practice (Bengio, Simard, & Frasconi, 1994). The classical Back Propagation Through Time learning algorithm is converging very slowly and the prediction performance sometimes deteriorates quickly when the network crosses a bifurcation.

Echo State Networks (ESN) are a special kind of recurrent neural networks designed for performing non-linear time-series forecasting (Jaeger, 2001; Jaeger & Haas, 2004). As an instance of a more general framework called reservoir computing (Lukosevicius & Jaeger, 2009), 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 (Buonomano & Merzenich, 1995), 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 (Galtier, Marini, Wainrib, & Jaeger, 2014), for which the associated gradient descent may be implemented with biologically plausible learning rules (Galtier & Wainrib, 2013). In this paper, we focus on the input-driven reservoir, which may be governed by a variety of dynamical systems beyond random neural networks (Dambre, Verstraeten, Schrauwen, & Massar, 2012), although we will only deal with RNN here.

It is important that the driven reservoir produces a trajectory robust to small perturbations. When this condition is unsatisfied, two arbitrarily close input signals may lead to two very different reservoir representations, hence making ESNs useless for supervised learning tasks. Originally, Jaeger has introduced the *Echo State Property* (ESP) which guarantees that the network is in a suitable state to do predictions (Jaeger, 2001).

Finding the right set of hyper-parameters, such as the spectral radius of the reservoir connectivity matrix, to achieve optimal performance is usually done through computationally intensive cross-validation. Indeed, the theoretical analysis of this question remains an open problem. Estimating the domain of validity of the ESP is the first step in this process, since it is a necessary condition to have a decent performance. To date, existing theoretical results, reviewed in Section 2, do not enable a practical estimation of this domain.

* Corresponding author.
E-mail address: gilles.wainrib@ens.fr (G. Wainrib).

http://dx.doi.org/10.1016/j.neunet.2015.12.013
0893-6080/

In this paper, we propose a practical algorithm for computing the domain of validity of the ESP. To achieve the goal, we use 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 unraveling 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 Massar and Massar (2013), this theoretical approach has a long history, dating back to early works on spin-glass models (Sompolinsky & Zippelius, 1981, 1982), followed by applications to random neural networks dynamics as in Cessac, Doyon, Quoy, and Samuelides (1994), Faugeras, Touboul, and Cessac (2009), Molgedey, Schuchhardt, and Schuster (1992) and Sompolinsky, Crisanti, and Sommers (1988). The rigorous justification of this heuristic approach is non-trivial and has been resolved by Arous and Guionnet (1995), Cabana and Touboul (2013) and Moynot and Samuelides (2002) using large deviations techniques.

The paper is organized as follows. In Section 2, we review existing theoretical results about the ESP. Next, in Section 3, 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. Finally, in Section 4 we show how this can be used to define a computable condition guaranteeing an operational version of the ESP.

2. Echo-state property: preliminary results

2.1. Echo-state network

The network we consider in this paper is a leaky integrator ESN (Jaeger, Lukosevicius, Popovici, & Siewert, 2007) 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 $u : \{1 \cdots T\} \to \mathbb{R}$, the classical neural network discrete dynamics is

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

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 $S(.)$ 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: $S(x) = \mathrm{erf}(\frac{\sqrt{\pi}}{2}x)$ (which follows the requirements above). We write $\sum_{j \to i}$ 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.

2.2. Echo-state property

The original definition of the ESP from Jaeger (2001) and the equivalent formulations manipulate left infinite input time-series assuming that the initial condition occurs at $t = -\infty$. The ESP definition can be summarized as

Definition 2.1 (*ESP Jaeger, 2001*)**.** 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 (Cheban, 2004). 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 $\tau = l = 1$, then the following result holds:

Theorem 2.1 (*Jaeger, 2001*)**.** *If $\eta < 1$, then the global ESP holds for every input.*

It is important to observe that the sufficient condition in 2.1 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 Zhang, Miller, and Wang (2012), 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 (Tao, Vu, & Krishnapur, 2010), 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 (Bai & Silverstein, 2010; Geman, 1980; Marchenko & Pastur, 1967) to show that $\eta \to 2\sigma$ when the number of neurons tends to infinity. From this result, as mentioned in Zhang et al. (2012), it is clear that the condition on the singular values translates to

Theorem 2.2. *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 (Lukosevisius, 2012). Based on the notion of structured singular value and on concepts from control theory (Lohmiller & Slotine, 1998), 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 (Buehner & Young, 2006). 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 (Zhang et al., 2012), where the concentration of measure phenomenon (Ledoux, 2005) is used to prove that: