Neurocomputing 298 (2018) 4–8

Contents lists available at ScienceDirect

Neurocomputing

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

Asymptotic Fisher memory of randomized linear symmetric Echo State Networks

Peter Tiňo

School of Computer Science, The University of Birmingham, Birmingham B15 2TT, United Kingdom

ARTICLE INFO

Article history:
Received 14 July 2017
Revised 22 October 2017
Accepted 26 November 2017
Available online 21 February 2018

Keywords:
Fisher memory of dynamical systems
Recurrent neural network
Echo State Network
Reservoir Computing

ABSTRACT

We study asymptotic properties of Fisher memory of linear Echo State Networks with randomized symmetric state space coupling. In particular, two reservoir constructions are considered: (1) More direct dynamic coupling construction using a class of Wigner matrices and (2) positive semi-definite dynamic coupling obtained as a product of unconstrained stochastic matrices. We show that the maximal Fisher memory is achieved when the input-to-state coupling is collinear with the dominant eigenvector of the reservoir coupling matrix. In the case of Wigner reservoirs we show that as the system size grows, the contribution to the Fisher memory of self-coupling of reservoir units is negligible. We also prove that when the input-to-state coupling is collinear with the sum of eigenvectors of the state space coupling, the expected normalized memory is four and eight time smaller than the maximal memory value for the Wigner and product constructions, respectively.

1. Introduction

Input driven dynamical systems play a prominent role in machine learning as models applied to time series data, e.g. [2,10,15,21]. There has been a lively research activity on formulating and assessing different aspects of computational power and information processing in such systems (see e.g. [5,16]). For example, tools of information theory have been used to assess information storage or transfer within systems of this kind [3,13,14,17]. Alternatively, dynamical systems have been assessed as feature generators for machine learning algorithms in terms of class separability (in sequence classification problems) or learnability [12].

To specifically characterize capability of input-driven dynamical systems to keep in their state-space information about past inputs, several memory quantifiers were proposed, for example *short term memory capacity* [9] and *Fisher memory curve* [6]. Even though those two measures have been developed from completely different perspectives, deep connections exist between them [20]. The concept of memory capacity, originally developed for univariate input streams, was generalized to multivariate inputs in [8]. Couillet et al. [4] rigorously studied mean-square error of linear dynamical systems used as dynamical filters in regression tasks and suggested memory quantities that generalize the short term memory capacity and Fisher memory curve measures. Finally, Ganguli and Sompolinski [7] showed an interesting connection between memory in dynamical systems and their capacity to perform dynamical compressed sensing of past inputs.

In this contribution we concentrate on Fisher memory of linear dynamical systems with symmetric coupling. In Echo State Networks (ESN) [15] large state space dimensionalities with random dynamical couplings are typically used and linear readout from the state space forms the only trainable part of the model. It is therefore important to characterize important large scale properties of Fisher memory in such systems (as the state space dimensionality grows) and study optimal settings of input-to-state couplings that maximize the memory. In particular, we rigorously study Fisher memory of two subclasses of linear input driven dynamical systems with symmetric dynamical coupling - a direct dynamic coupling construction using a class of Wigner matrices (Section 3) and a positive semi-definite dynamic coupling obtained as a product of unconstrained stochastic matrices (Section 4).

2. Fisher memory curve of linear dynamical systems

We consider linear input driven dynamical systems with N-dimensional state space and univariate inputs and outputs with randomized symmetric dynamic coupling.

In the ESN metaphor, the state dimensions correspond to reservoir units coupled to the input $s(t)$ and output $y(t)$ through N-dimensional weight vectors $\mathbf{v} \in \mathbb{R}^N$ and $\mathbf{r} \in \mathbb{R}^N$, respectively. Denoting the state vector at time t by $\mathbf{x}(t) \in \mathbb{R}^N$, the dynamical system

E-mail address: p.tino@cs.bham.ac.uk

https://doi.org/10.1016/j.neucom.2017.11.076
0925-2312/

(reservoir activations) evolves as

$$\mathbf{x}(t) = \mathbf{v}s(t) + \mathbf{W}\mathbf{x}(t-1) + \mathbf{z}(t), \quad (1)$$

where $\mathbf{W} \in \mathbb{R}^{N\times N}$ is a $N \times N$ weight matrix providing the dynamical coupling and $\mathbf{z}(t)$ are zero-mean noise terms. Parameters $\mathbf{r}$ of the adaptive linear readout, $y(t) = \mathbf{r}^T\mathbf{x}(t)$, are typically trained (offline or online) by minimizing the (normalized) mean square error between the targets and reservoir readouts $y(t)$. For our analysis, however, the readout part of the ESN architecture is not needed.

In ESN, the elements of $\mathbf{W}$ and $\mathbf{v}$ are fixed prior to training, often at random, with entries drawn from a distribution symmetric with respect to the origin. The reservoir connection matrix $\mathbf{W}$ is typically scaled to a prescribed spectral radius < 1, although in this study we assume that the parameters of the distribution over $\mathbf{W}$ are set so that asymptotically, almost surely, $\mathbf{W}$ is a contractive linear operator.

In [6] Ganguli et al. proposed a particular way of quantifying the amount of memory preserved in linear input driven dynamical systems corrupted by a memoryless Gaussian i.i.d dynamic noise[1] $\mathbf{z}(t)$. In particular, $\mathbf{z}(t)$ is zero mean with co-variance $\epsilon\mathbf{I}$, $\epsilon > 0$, where $\mathbf{I}$ is the $N \times N$ identity matrix. Under such dynamic noise, given an input driving stream $s(..t) = ... \; s(t-2) \; s(t-1) \; s(t)$, the input-conditional state distribution

$$p(\mathbf{x}(t)|\; ... \; s(t-2) \; s(t-1) \; s(t))$$

is a Gaussian with covariance [6]

$$\mathbf{C} = \epsilon \sum_{\ell=0}^{\infty} \mathbf{W}^{\ell}(\mathbf{W}^T)^{\ell}. \quad (2)$$

Sensitivities of $p(\mathbf{x}(t)|\, s(..t))$ with respect to small perturbations in the input driving stream $s(..t)$ (parameters of the dynamical system remain fixed) are collected in the Fisher memory matrix $\mathbf{F}$ with elements

$$F_{k,l}(s(..t)) = -\mathbb{E}_{p(x(t)|\; s(..t))}\left[\frac{\partial^2}{\partial s(t-k)\partial s(t-l)} \log p(\mathbf{x}(t)|\; s(..t))\right]$$

and its diagonal elements $J_N(k) = F_{k,k}(s(..t))$ quantify the information that the state distribution $p(x(t)|\, s(..t))$ retains about a change (e.g. a pulse) entering the network $k > 0$ time steps in the past. The collection of terms $\{J_N(k)\}_{k=0}^{\infty}$ was termed Fisher memory curve (FMC) and evaluated to [6]

$$J_N(k; \mathbf{W}, \mathbf{v}) = \mathbf{v}^T(\mathbf{W}^T)^k\mathbf{C}^{-1}\mathbf{W}^k\mathbf{v}, \quad (3)$$

where in the notation $J_N(k; \mathbf{W}, \mathbf{v})$ we made explicit the dependence of FMC on the dynamic and input and couplings $\mathbf{W}$ and $\mathbf{v}$, respectively.

Analogously to memory capacity of dynamical systems [9], we extend the Fisher memory curve to the global memory quantification,

$$\mathcal{J}_N(\mathbf{W}_N, \mathbf{v}) = \sum_{k=1}^{\infty} J_N(k; \mathbf{W}_N, \mathbf{v}).$$

We will refer to $\mathcal{J}_N(\mathbf{W}_N, \mathbf{v})$ as Fisher memory of the underlying dynamical system. Obviously, increasing state space dimension N will increase the amount of memory that can be usefully captured by the dynamical system (1). To remove this bias, we introduce a new quantity, *normalized Fisher memory*, which measures the amount of memory realisable by the dynamical system *per state space dimension*:

$$\bar{\mathcal{J}}_N(\mathbf{W}_N, \mathbf{v}) = \frac{1}{N}\mathcal{J}_N(\mathbf{W}_N, \mathbf{v}).$$

In the following we study asymptotic properties of the normalized Fisher memory as the state space dimensionality grows and ask what kind of input coupling $\mathbf{v}$ is needed to maximize its expectation. Again, it is important to realize that as the state space dimensionality N grows, so does the input weight dimensionality. Keeping the input weight norm constant while increasing the state space dimensionality would result in diminishing individual weights. To normalize the scales, so that asymptotic statements can be made, we will require that the input weights live on $(N-1)$-dimensional hypersphere, $\mathbf{v} \in S_{N-1}(\sqrt{N})$, where for $r > 0$,

$$S_{N-1}(r) = \{\mathbf{v} \in \mathbb{R}^N |\; \|\mathbf{v}\|_2 = r\}.$$

3. Wigner ESN

Theory of random matrices has undergone considerable development, see e.g. [19]. In this contribution we will study dynamical systems with randomized coupling constrained to the class of Wigner matrices (e.g. [1]). Let $\mathbf{Q}_N$ be a random symmetric $N \times N$ matrix with "upper triangular" off-diagonal elements $Q_{i,j}$, $1 \le i < j \le N$ distributed i.i.d. with zero mean and finite moments - in particular, of variance $\sigma_o^2 > 0$. Diagonal elements $Q_{i,i}$, $1 \le i \le N$ of $\mathbf{Q}_N$ are distributed i.i.d. with a zero-mean distribution of finite moments and variance $\sigma_d^2 > 0$. The elements below the diagonal are copies of their symmetric counterparts: for $1 \le j < i \le N$, $Q_{i,j} = Q_{j,i}$. Asymptotic properties of such matrices have been intensively studied, in particular the convergence of eigenvalues, as $N \to \infty$. It can be shown that in the general case, scaling down of random matrices is necessary to ensure convergence of their spectral properties [1]:

$$\mathbf{W}_N = \frac{1}{\sqrt{N}}\mathbf{Q}_N.$$

We will refer to ESN with dynamical coupling $\mathbf{W}_N$ as Wigner Echo State Networks. We are now ready to state the first result concerning maximal Fisher memory of Wigner ESNs.

Theorem 1. *Consider a sequence of Wigner dynamical systems (1) with couplings $\{\mathbf{W}_N\}_{N>1}$. The maximum normalized Fisher memory is attained when for every realization of Wigner coupling $\mathbf{W}_N$, the input weights $\mathbf{v}$ are collinear with the dominant eigenvector[2] of $\mathbf{W}_N$. In that case, as $N \to \infty$, almost surely,*

$$\bar{\mathcal{J}}_N(\mathbf{W}_N, \mathbf{v}) \to \frac{4}{\epsilon}\,\sigma_o^2.$$

Proof. For a fixed N, let $\mathbf{W}_N$ be a realization of Wigner coupling. Since $\mathbf{W}_N$ is symmetric, it can be diagonalised,

$$\mathbf{W}_N = \mathbf{U}_N\Lambda_N\mathbf{U}_N^T, \quad \Lambda_N = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N). \quad (4)$$

Without loss of generality assume $\lambda_1 \ge \lambda_2 \ge \cdots \lambda_N$. Columns of $\mathbf{U}_N$ are eigenvectors $\{\mathbf{u}_i\}_{i=1}^N$ of $\mathbf{W}_N$, forming an orthonormal basis of $\mathbb{R}^N$. Let $\tilde{\mathbf{v}}$ be the expression of input weights $\mathbf{v}$ in this basis, i.e. $\tilde{\mathbf{v}} = \mathbf{U}_N^T\mathbf{v}$. It has been shown in [20] that for symmetric dynamic couplings,

$$J_N(k; \mathbf{W}_N, \mathbf{v}) = \frac{1}{\epsilon}\sum_{i=1}^{N} \tilde{v}_i^2\;\lambda_i^{2k}\;(1-\lambda_i^2).$$

[1] As customary in the dynamical systems literature, we distinguish between the "observational" and "dynamic" noise. Observational noise refers to the noise applied to readouts from the state space in the process of their measurement. This noise does not corrupt the underlying dynamics of the system. On the other hand, dynamic noise corrupts the system dynamics in the state space. The term dynamic noise does not in this case refer to the possibility of its distribution changing in time.

[2] the eigenvector corresponding to the maximal eigenvalue.