ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

Fisher Memory of Linear Wigner Echo State Networks

Peter Tiňo[1] *

School of Computer Science, The University of Birmingham
Birmingham B15 2TT, United Kingdom
P.Tino@cs.bham.ac.uk

Abstract. We study asymptotic properties of Fisher memory of linear Echo State Networks with randomized reservoir coupling prescribed by the class of Wigner matrices. Several properties of Fisher memory normalized per state space dimension are derived. In particular, we show that as the system size grows, the contribution of self-coupling of self-loops on reservoir units to the Fisher memory is negligible. Furthermore, we prove that the maximal Fisher memory is achieved when the input-to-state coupling is colinear with the dominant eigenvector of the state space coupling matrix. Finally, we show that when the input-to-state coupling is colinear with the sum of eigenvectors of the state space coupling, the expected normalized memory is four time smaller than the maximal memory value.

1 Introduction

Input driven dynamical systems play a prominent role in machine learning as models applied to time series data, e.g. [1, 2, 3, 4]. To characterize dynamic properties of such systems, several memory quantifiers of such systems were proposed (e.g. [5, 6, 7]), in particular the *Fisher memory curve* $J(k)$ [8]. In Echo State Networks (ESN) [4] 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 asymptotic 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 this contribution we rigorously study Fisher memory of a subclass of linear input driven dynamical systems whose dynamical coupling is formed by Wigner random matrices. In particular, we derive closed form expressions for the maximum Fisher memory that can be attained by such systems and show that the role of self-coupling in Fisher memory of Wigner dynamical systems is vanishingly negligible as the system size grows.

2 Background

We consider linear input driven dynamical systems with N-dimensional state space and univariate inputs and outputs with randomized dynamic coupling prescribed by the class of Wigner matrices.

*This work was supported by EPSRC grant EP/L000296/1.

In the (linear) 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 $\mathbf{u} \in \mathbb{R}^N$, respectively. Denoting the state vector at time t by $\mathbf{x}(t) \in \mathbb{R}^N$, the dynamical system (reservoir activations) evolve as

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

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.

2.1 Fisher memory curve

In [8] Ganguli, Huh and Sompolinsky 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 $\mathbf{z}(t)$. 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 [8]

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

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 [8]

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

where in the notation we made explicit the dependence on the dynamic and input and couplings $\mathbf{W}$ and $\mathbf{v}$, respectively.

2.2 Wigner ESN

Theory of random matrices has undergone considerable development [9]. In this contribution we will study dynamical systems with randomized coupling constrained to the class of Wigner matrices [10]. Let $\mathbf{X}_N$ be a random symmetric $N \times N$ matrix with "upper triangular" off-diagonal elements $X_{i,j}$, $1 \leq i < j \leq N$ distributed i.i.d. with zero mean and finite moments - in particular, of variance $\sigma_o^2 > 0$. Diagonal elements $X_{i,i}$, $1 \leq i \leq N$ of $\mathbf{X}$ 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 $1 \leq j < i \leq N$, $X_{i,j} = X_{j,i}$. 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 [10]:

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

Matrices $\mathbf{W}_N \in \mathbb{R}^{N \times N}$ represent a special class of Wigner matrices and we will refer to linear ESN with dynamical coupling $\mathbf{W}_N$ as Wigner Echo State Networks.

3 Fisher memory of Wigner ESN

As in the memory capacity of dynamical systems [11], the Fisher memory curve can also be extended to the global memory quantification [8],

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

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 a 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 will ask what kind of input coupling $\mathbf{v}$ is needed to maximize its expectation. Again, it is important to realise that as the state space dimensionality N grows, so does the input weight dimensionality. In order to normalise 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 $r > 0$,

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

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 realisation of Wigner coupling* $\mathbf{W}_N$*, the input weights* $\mathbf{v}$ *are colinear with the dominant eigenvector*[1] *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 realisation of Wigner coupling. Since $\mathbf{W}_N$ is symmetric, it can be diagonalised,

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

Without loss of generality assume $\lambda_1 \geq \lambda_2 \geq ...\lambda_N$. 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 [12] that for symmetric dynamic couplings,

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

We therefore have

$$\begin{aligned}\epsilon \cdot \mathcal{J}_N(\mathbf{W}_N, \mathbf{v}) &= \sum_{k=1}^\infty \sum_{i=1}^N \tilde{v}_i^2\ \lambda_i^{2k}\ (1 - \lambda_i^2) = \sum_{i=1}^N \tilde{v}_i^2\ (1 - \lambda_i^2) \sum_{k=1}^\infty \lambda_i^{2k} \\ &= \sum_{i=1}^N \tilde{v}_i^2\ \lambda_i^2. \end{aligned} \tag{5}$$

Letting $N^{-1/2}\mathbf{v}$ be the dominant eigenvector $\mathbf{u}_1$ of $\mathbf{W}_N$ results in

$$\tilde{\mathbf{v}} = \sqrt{N}\ \mathbf{U}_N^T \mathbf{u}_1 = \sqrt{N}\ \mathbf{e}_1,$$

where $\mathbf{e}_i$ the i-th standard basis vector, i.e. $\mathbf{e}_i \in \mathbb{R}^N$ is the vector of 0's, except for the value 1 at index i. We thus have

$$\epsilon \cdot \mathcal{J}_N(\mathbf{W}_N, \mathbf{u}_1) = N\ \lambda_1^2$$

and hence $\epsilon \cdot \bar{\mathcal{J}}_N(\mathbf{W}_N, \mathbf{u}_1) = \lambda_1^2$. Now, the maximal eigenvalue of Wigner matrices converges to $2\sigma_o$ almost surely [10], giving the almost sure convergence of $\bar{\mathcal{J}}_N(\mathbf{W}_N, \mathbf{u}_1)$ to $(4\sigma_o^2)/\epsilon$.

To show that colinearity of input weights $\mathbf{v} \in S_{N-1}(\sqrt{N})$ with dominant eigenvector of $\mathbf{W}_N$ is the optimal setting, we note that since $\tilde{\mathbf{v}}$ expresses $\mathbf{v}$ in another orthonormal basis $\mathbf{U}_N$, the norm is preserved, $\|\mathbf{v}\|_2 = \|\tilde{\mathbf{v}}\|_2$. Hence, $\tilde{\mathbf{v}} \in$

[1] the eigenvector corresponding to the maximal eigenvalue

$S_{N-1}(\sqrt{N})$. In line with eq. (5) we therefore consider the following optimization problem:

$$\max_{\mathbf{q}\in S_{N-1}(1)} N \cdot \sum_{i=1}^{N} q_i^2 \ \lambda_i^2.$$

Reparametrization $a_i = q_i^2$, $b_i = \lambda_i^2$ and ignoring constant scaling leads to

$$\max_{\mathbf{a}\in \mathcal{Q}_{N-1}} \mathbf{b}^T\mathbf{a}, \tag{6}$$

which is a linear optimization problem on simplex

$$\mathcal{Q}_{N-1} = \{\mathbf{x} \in [0.1]^N | \ \|\mathbf{x}\|_1 = 1\}.$$

Let the largest element of $\mathbf{b}$ be b_{i_*}, i.e. $i_* = \arg\max_i b_i$. Then the quantity in (6) is maximised when $\mathbf{a} = \mathbf{e}_{i_*}$, in other words, $a_{i_*} = 1$ and $a_j = 0$, $j \neq i_*$. In our case $i_* = 1$ and so the non-zero element of $\mathbf{a}$ is $a_1 = q_1 = 1$. It follows that $\tilde{v}_1^2 = N$ and $\tilde{v}_j = 0$, $j = 2, 3, ..., N$. This directly implies $\mathbf{v} \in S_{N-1}$ and colinear with $\mathbf{u}_1$. □

The last result imposes an asymptotic upper bound on normalized Fisher memory of Wigner ESNs. Obviously, $\bar{\mathcal{J}}_N(\mathbf{W}_N, \mathbf{v})$ can be made vanishingly small by making the input weights $\mathbf{v}$ colinear with the least significant eigenvector of $\mathbf{W}_N$ (see semicircular law of eigenvalue distribution for Wigner matrices [10]. But what about other cases, e.g. when the inputs weights $\mathbf{v}$ are colinear with the sum of eigenvectors of $\mathbf{W}_N$?

Theorem 2: *Consider a sequence of Wigner dynamical systems* (1) *with couplings* $\{\mathbf{W}_N\}_{N>1}$. *For every realisation of Wigner coupling* $\mathbf{W}_N$, *let the input weights* $\mathbf{v}$ *be colinear with the sum of eigenvectors of* $\mathbf{W}_N$. *Then, as* $N \to \infty$, *for the expected normalized Fisher memory we have,*

$$\mathbb{E}_{\mathbf{W}_N}[\bar{\mathcal{J}}_N(\mathbf{W}_N, \mathbf{v})] \to \frac{1}{\epsilon} \ \sigma_o^2.$$

<u>Proof:</u> In this case, $\mathbf{v} = \sum_{i=1}^{N} \mathbf{u}_i \in S_{N-1}(\sqrt{N})$, $\tilde{v}_i = 1$, $i = 1, 2, ..., N$. By (5),

$$\|\mathbf{W}_N\|_F^2 = \sum_{i,j=1}^{N} W_{N,ij}^2 = \sum_{i=1}^{N} \lambda_i^2 = \epsilon \cdot \mathcal{J}_N(\mathbf{W}_N, \mathbf{v}),$$

where $\|\cdot\|_F$ is the Frobenius norm. This implies

$$\begin{aligned} \epsilon \cdot \mathbb{E}_{\mathbf{W}_N}[\bar{\mathcal{J}}_N(\mathbf{W}_N, \mathbf{v})] &= \frac{1}{N} \sum_{i,j=1}^{N} \mathbb{E}[W_{N,ij}^2] \\ &= \frac{1}{N}\sigma_d^2 + \frac{N-1}{N}\sigma_o^2 \end{aligned}$$

and the result follows from sending $N \to \infty$. □

4 Conclusion

We have rigorously studied Fisher memory of a subclass of linear input driven dynamical systems whose dynamical coupling is formed by Wigner random matrices. Such systems can be viewed as Echo State Networks with Wigner reservoir coupling. In order to study how memory properties of such systems scale with the system size, we investigated Fisher memory normalized per state space dimension. Several interesting findings were derived, in particular: (1) as the system size grows, the contribution of self-coupling of the states (self-loops on reservoir units in ESN) to the normalized Fisher memory is negligible; (2) the maximal normalized Fisher memory is achieved when the input-to-state coupling is colinear with the dominant eigenvector of the state space coupling matrix; and (3) when the input-to-state coupling is colinear with the sum of eigenvectors of the state space coupling, the expected normalized memory is four time smaller than the maximal memory value achieved when colinearity with the dominant eigenvector only is employed.

References

[1] Witali Aswolinskiy, Felix Reinhart, and Jochen J. Steil. Modelling of Parameterized Processes via Regression in the Model Space. In *Proceedings of 24th European Symposium on Artificial Neural Networks*, pages 53–58, 2016.

[2] H. Jaeger and H. Hass. Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless telecommunication. *Science*, 304:78–80, 2004.

[3] M. H. Tong, A.D. Bicket, E.M. Christiansen, and G.W. Cottrell. Learning grammatical structure with echo state network. *Neural Networks*, 20:424–432, 2007.

[4] M. Lukosevicius and H. Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149, 2009.

[5] T. Toyoizumi. Nearly extensive sequential memory lifetime achieved by coupled nonlinear neurons. *Neural Computation*, 24:2678–2699, 2012.

[6] A.S. Charles, H.L. Yap, and Ch.J. Rozell. Short-term memory capacity in networks via the restricted isometry property. *Neural Computation*, 26:1198–1235, 2014.

[7] B. Zhang, D.J Miller, and Y. Wang. Nonlinear system modeling with random matrices: Echo state networks revisited. *IEEE Transactions on Neural Networks and Learning Systems*, 23(1):175–182, 2007.

[8] S. Ganguli, D. Huh, and H. Sompolinsky. Memory traces in dynamical systems. *Proceedings of the National Academy of Sciences*, 105:18970–18975, 2008.

[9] T. Tao. *Topics in Random Matrix Theory.* American Mathematical Society, Graduate Studies in Mathematics, 2012.

[10] G. Anderson, A. Guionnet, and O. Zeitouni. *An Introduction to Random Matrices (Cambridge Studies in Advanced Mathematics).* Cambridge University Press, 2010.

[11] H. Jaeger. Short term memory in echo state networks. Technical report gmd report 152, German National Research Center for Information Technology, 2002.

[12] P. Tiňo and A. Rodan. Short term memory in input-driven linear dynamical systems. *Neurocomputing*, 112:58–63, 2013.