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
E-mail: P.Tino@cs.bham.ac.uk

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.

Keywords: Fisher memory of dynamical systems, Recurrent neural network, Echo state network, Reservoir Computing

1. Introduction

Input driven dynamical systems play a prominent role in machine learning as models applied to time series data, e.g. [2, 10, 21, 15]. 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 [13, 14, 17, 3].

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 ons 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 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 (reservoir activations) evolves 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.

In [6] 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[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]

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

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

[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.

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]

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

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 $Q_{i,j}$, $1 \leq i < j \leq N$ distributed i.i.d. with zero mean and finite moments - in particular, of variance $\sigma_o^2 > 0$. Diagonal elements $Q_{i,i}$, $1 \leq i \leq N$ 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 $1 \leq j < i \leq N$, $Q_{i,j} = Q_{j,i}$. 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,

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

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

[2] the eigenvector corresponding to the maximal eigenvalue

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,

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

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 [1], 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 collinearity 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, $\tilde{\mathbf{v}} \in S_{N-1}(\sqrt{N})$. 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

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

which is a linear optimization problem on simplex

$$\mathcal{Z}_{N-1} = \{\mathbf{a} \in [0,1]^N |\ \|\mathbf{a}\|_1 = 1\}.$$

Let the largest element of $\mathbf{b}$ be b_{i_*}, i.e. $i_* = \arg\max_i b_i$. Then the quantity in (6) is maximized 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 collinear 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 $\mathbf{v}$ collinear with the least significant eigenvector of $\mathbf{W}_N$ (see semicircular law of eigenvalue distribution for Wigner matrices [1]). In the following we ask to what degree will the Fisher memory of (1) degrade if instead of the dominant eigenvector the input weights are made collinear 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 realization of Wigner coupling* $\mathbf{W}_N$, *let the input weights* $\mathbf{v}$ *be collinear with the sum of eigenvectors of* $\mathbf{W}_N$. *Then, as* $N \to \infty$, *for the expected normalized Fisher memory we have,*

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

Proof: 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),

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

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. Symmetric Positive Semi-definite Reservoirs

In the previous section, the symmetry of randomized dynamic reservoirs was imposed directly by stipulating that the matrix elements $Q_{i,j}$ and $Q_{j,i}$ must be equal. This is achieved by randomly generating $Q_{i,j}$ above the diagonal and then copying $Q_{i,j}$ to their symmetric counterparts $Q_{j,i}$ below the diagonal (Wigner random matrices). In this section we ask whether the nature of the results derived for Wigner ESN changes if another, less direct construction of randomized symmetric reservoirs is used. In particular, we generate $\mathbf{Q}_N$ as

$$\mathbf{Q}_N = \mathbf{Y}_N^T\mathbf{Y}_N, \tag{7}$$

where $\mathbf{Y}_N$ is a random $N \times N$ matrix with elements $Y_{i,j}$, $1 \leq i, j \leq N$, distributed i.i.d. with zero mean and finite moments - in particular, of variance $\sigma_Y^2 > 0$. As before, to ensure convergence of spectral properties of random matrices $\mathbf{Y}_N$, a scaling factor of $N^{-1/2}$ needs to be applied to $\mathbf{Y}_N$ [1], resulting in:

$$\mathbf{W}_N = \frac{1}{N}\mathbf{Q}_N. \tag{8}$$

Note that each realization of $\mathbf{Y}_N$ yields a positive semi-definite matrices $\mathbf{Q}_N$ and $\mathbf{W}_N$.

Recall that to show that the optimal (i.e. maximizing Fisher memory) setting of input weights (up to necessary scaling) is the leading eigenvector of $\mathbf{W}_N$, we only needed symmetry of $\mathbf{W}_N$. Considering that the individual items of random matrix are generated i.i.d. from a 0-mean distribution with variance σ_Y^2, the question is to what extent does the particular randomized

construction of $\mathbf{W}_N$ matter when asymptotic properties of Fisher memory are considered. We have the following result:

Theorem 3: *Consider a sequence of dynamical systems* (1) *with randomized positive semidefinite couplings* $\{\mathbf{W}_N\}_{N>1}$ (7)-(8). *If the input weights* $\mathbf{v}$ *are collinear with the dominant eigenvector of* $\mathbf{W}_N$*, as* $N \to \infty$*, almost surely,*

$$\bar{\mathcal{J}}_N(\mathbf{W}_N, \mathbf{v}) \to \frac{16}{\epsilon}\, \sigma_Y^4.$$

Proof: Let $s_{max}(\mathbf{Y}_N)$ be the maximal singular value of the $N \times N$ matrix $\mathbf{Y}_N$. As $N \to \infty$,

$$\frac{1}{\sqrt{N}} s_{max}(\mathbf{Y}_N) \to 2\sigma_Y$$

almost surely [18]. This implies that the maximum eigenvalue $\lambda_{max}(\mathbf{W}_N)$ of $\mathbf{W}_N$ approaches $4\sigma_Y^2$.

We have shown in the proof of Theorem 1 that if the input weight is collinear with the dominant eigenvector of $\mathbf{W}_N$, we have $\epsilon \cdot \bar{\mathcal{J}}_N = \lambda_{max}^2(\mathbf{W}_N)$. Hence, as $N \to \infty$, $\bar{\mathcal{J}}_N(\mathbf{W}_N, \mathbf{v})$ converges to $16\ \sigma_Y^4/\epsilon$ almost surely. □

Theorem 3 demonstrates that the asymptotic properties of Fisher memory translate directly to the case of symmetric randomized dynamical couplings constructed as products of unconstrained random matrices. Of course, since $\mathbf{W}_N$ is now constructed as a *product* of random matrices, it is natural that the asymptotic Fisher memory is expressed as a square of the expression in Theorem 1. Indeed, consider an alternative construction of the symmetric reservoir:

$$\tilde{\mathbf{Q}}_N = \left[\mathbf{Y}_N^T \mathbf{Y}_N\right]^{\frac{1}{2}}. \tag{9}$$

Each realization of $\mathbf{Y}_N$ yields a positive semi-definite matrix $\mathbf{Y}_N^T \mathbf{Y}_N$ that has a unique square root - a realization of $\tilde{\mathbf{Q}}_N$. Again, to ensure convergence of spectral properties of random matrices $\mathbf{Y}_N$, a scaling factor of $N^{-1/2}$ needs to be applied:

$$\tilde{\mathbf{W}}_N = \frac{1}{\sqrt{N}} \tilde{\mathbf{Q}}_N, \tag{10}$$

Now, since $N^{-1/2} s_{max}(\mathbf{Y}_N) \to 2\sigma_Y$ almost surely, as $N \to \infty$,

$$\frac{1}{N} \lambda_{max}(\mathbf{Y}_N^T \mathbf{Y}_N) \to 4\sigma_Y^2.$$

Hence, the maximum eigenvalue $\lambda_{max}(\tilde{\mathbf{W}}_N)$ of $\tilde{\mathbf{W}}_N$ approaches $2\sigma_Y$. It follows that if the input weight is collinear with the dominant eigenvector of $\tilde{\mathbf{W}}_N$, we have $\epsilon \cdot \bar{\mathcal{J}}_N = \lambda^2_{max}(\tilde{\mathbf{W}}_N)$ and as $N \to \infty$, $\bar{\mathcal{J}}_N(\tilde{\mathbf{W}}_N, \mathbf{v})$ converges to $4\ \sigma_Y^2/\epsilon$ almost surely. This is in direct correspondence with the result of Theorem 1 for Wigner ESN. In this case, irrespective of whether the symmetry of randomized dynamical coupling is obtained directly (Wigner ESN), or indirectly as a product of unconstrained random matrices, the maximal normalized Fisher memory approaches $\frac{4}{\epsilon}\ \sigma^2$ and is determined solely by the second moment of the zero-mean random variables generating the strength of dynamic couplings within the reservoir.

We now turn our attention to the case of input weight vector collinear with the sum of eigenvectors of $\mathbf{W}_N$. We have shown that, compared to the maximum Fisher memory (input weight vector collinear with the leading eigenvector of $\mathbf{W}_N$), this results in reduced asymptotic Fisher memory by a factor of 4 (see Theorems 1 and 2). Is this a property of the particular randomized construction of symmetric dynamic coupling as a Wigner matrix, or does it reflect a more general tendency?

Theorem 4: *Consider a sequence of Wigner dynamical systems* (1) *with couplings* $\{\mathbf{W}_N\}_{N>1}$ (7)-(8). *Assume that the entries of* $\mathbf{Y}_N$ *are i.i.d. generated from a symmetric distribution with 0-mean, variance* σ_Y^2 *and finite first four moments. For every realization of dynamic coupling* $\mathbf{W}_N$*, let the input weights* $\mathbf{v}$ *be collinear with the sum of eigenvectors of* $\mathbf{W}_N$*. Then, as* $N \to \infty$*, for the expected normalized Fisher memory we have,*

$$\mathbb{E}_{\mathbf{W}_N}[\bar{\mathcal{J}}_N(\mathbf{W}_N, \mathbf{v})] \to \frac{2}{\epsilon}\ \sigma_Y^4.$$

Proof: In the proof of Theorem 2 we have shown that if the input weight vector $\mathbf{v}$ is collinear with the sum of eigenvectors of $\mathbf{W}_N$, then

$$\epsilon \cdot \mathcal{J}_N(\mathbf{W}_N, \mathbf{v}) = \sum_{i=1}^{N} \lambda_i^2, \tag{11}$$

where λ_i's are eigenvalues of $\mathbf{W}_N$. Furthermore,

$$\sum_{i=1}^{N} \lambda_i^2 = trace(\mathbf{W}_N^2), \tag{12}$$

and so the expected Fisher memory can be evaluated as

$$\mathbb{E}_{\mathbf{Y}_N}[\mathcal{J}_N(\mathbf{W}_N, \mathbf{v})] = \frac{1}{\epsilon}\mathbb{E}_{\mathbf{Y}_N}[trace(\mathbf{W}_N^2)],$$

where

$$\mathbf{W}_N^2 = \frac{1}{N^2}\mathbf{Y}_N^T\mathbf{Y}_N\mathbf{Y}_N^T\mathbf{Y}_N.$$

Fix a positive semi-definite $N \times N$ matrix Γ_N with eigenvalues $\gamma_1, ..., \gamma_N$. Kabán showed ([11], Lemma 2) that the expectation of $\mathbf{Y}_N^T\mathbf{Y}_N\Gamma_N\mathbf{Y}_N^T\mathbf{Y}_N$ reads

$$\mathbb{E}_{\mathbf{Y}_N}[\mathbf{Y}_N^T\mathbf{Y}_N\Gamma_N\mathbf{Y}_N^T\mathbf{Y}_N] = N\sigma_Y^4\left[(N+1)\Gamma_N + trace(\Gamma_N)\mathbf{I}_N + \mathcal{E}\sum_{i=1}^{N}\gamma_i\mathbf{A}^{(i)}\right],$$

where $\mathbf{I}_N$ is $N \times N$ identity matrix, $\mathcal{E}$ is excess kurtosis of the distribution generating elements of $\mathbf{Y}_N$ and $\mathbf{A}^{(i)}$ are $N \times N$ diagonal matrices with j-th diagonal elements equal to $\sum_{a=1}^{N} u_{a,i}^2 u_{a,j}^2$ and $u_{a,i}$ is the a-th item of the i-th eigenvector of Γ_N.

In our case $\Gamma_N = \mathbf{I}_N$ with eigenvectors $\mathbf{e}_i$ standard basis (all elements equal to 0 and the i-th element 1) and eigenvalues $\gamma_1 = \gamma_2 = ... = \gamma_N = 1$. Note that $(\mathbf{e}_i)_a = \delta(i, a)$, where δ is the Kronecker delta. Hence,

$$\mathbf{A}_{j,j}^{(i)} = \sum_{a=1}^{N}\delta(i,a)\delta(j,a) = \delta(i,j)$$

and so $\mathbf{A}^{(i)} = diag(\mathbf{e}_i)$. It follows that

$$\sum_{i=1}^{N}\gamma_i\mathbf{A}^{(i)} = \mathbf{I}_N.$$

We can now evaluate

$$\begin{aligned} N^2\,\mathbb{E}_{\mathbf{Y}_N}[\mathbf{W}_N^2] &= \mathbb{E}_{\mathbf{Y}_N}[\mathbf{Y}_N^T\mathbf{Y}_N\mathbf{Y}_N^T\mathbf{Y}_N] \\ &= \mathbb{E}_{\mathbf{Y}_N}[\mathbf{Y}_N^T\mathbf{Y}_N\mathbf{I}_N\mathbf{Y}_N^T\mathbf{Y}_N] \\ &= N\sigma_Y^4\left[(N+1)\mathbf{I}_N + N\mathbf{I}_N + \mathcal{E}\mathbf{I}_N\right] \\ &= N(2N+1+\mathcal{E})\sigma_Y^4\mathbf{I}_N. \end{aligned} \tag{13}$$

It follows that

$$\mathbb{E}_{\mathbf{Y}_N}[\mathbf{W}_N^2] = \sigma_Y^4 \left(2 + \frac{1+\mathcal{E}}{N}\right) \mathbf{I}_N.$$

We thus have

$$\mathbb{E}_{\mathbf{Y}_N}[trace(\mathbf{W}_N^2)] = \sigma_Y^4 \left(2N + 1 + \mathcal{E}\right)$$

From (11) and (12) we conclude that as $N \to \infty$, the normalized expected Fisher memory per dimension,

$$\bar{\mathcal{J}}_N(\mathbf{W}_N, \mathbf{v}) = \frac{1}{N \cdot \epsilon} \mathbb{E}_{\mathbf{Y}_N}[trace(\mathbf{W}_N^2)],$$

converges to $2\sigma_Y^4/\epsilon$. □

This result offers an interesting insight into the asymptotic behavior of Fisher memory for symmetric dynamic reservoirs. Whereas in the case of direct Wigner construction, when changing the input weight vector from the optimal setting of the leading eigenvector of dynamic coupling $\mathbf{W}_N$ to the sum of its eigenvectors, the Fisher memory drops by a factor of 4, in the case of symmetric positive semi-definite $\mathbf{W}_N$ obtained as a product of random matrices, the drop in Fisher memory is twice as large - by a factor of 8 (see Theorems 3 and 4). A deeper investigation of this phenomenon is beyond the scope of this study and here we can only speculate. It is possible that the faster drop in the Fisher memory reflects the fact that while the diagonal and upper-off-diagonal coupling weights in Wigner ESN are truly independent, in the positive semi-definite construction a more rigid dependency structure is imposed.

5. Discussion and Conclusion

We have rigorously studied Fisher memory of two subclasses of linear input driven dynamical systems. In order to study how memory properties of such systems scale with the system size, we investigated Fisher memory normalized per state space dimension.

The first subclass has a dynamical coupling formed by Wigner random matrices. Such systems can be viewed as Echo State Networks with Wigner reservoir coupling. 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 collinear with the dominant eigenvector of the state space coupling matrix; and
3. when the input-to-state coupling is collinear 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 collinearity with the dominant eigenvector only is employed.

Dynamical coupling of the second subclass of randomized symmetric reservoirs constructs positive semi-definite dynamical couplings as products of unconstrained random matrices. Interestingly enough, while in the case of Wigner reservoirs, when changing the input weight vector from the optimal setting of the leading eigenvector of dynamic coupling to the sum of its eigenvectors, the Fisher memory drops by a factor of 4, in the case of symmetric positive semi-definite couplings obtained as a product of random matrices, the drop in the Fisher memory is twice as large.

Note that in the case of positive semi-definite dynamical couplings, we no longer have the possibility of setting variances of self-loop weights independently of the variances of the inter-neuron connections. Hence, direct investigation of the influence of self-loops on Fisher memory in the asymptotic regime is not possible. However, one suspects that, as in the case of Wigner reservoirs, the main contributions to memory come from the inter-neuron couplings and for large reservoirs, memories with, or without self-loops will not differ significantly. Nevertheless, showing this rigorously will require considering more complex reservoir weight generation mechanisms that are beyond the scope of this study.

One can legitimately ask whether the restriction to linear and symmetric reservoirs does not severely limit the scope of this study with respect to reservoirs used in practice. First, in general, memory quantifications of dynamical systems, such as memory capacity or Fisher memory, quantify capabilities of dynamical systems that are not necessarily directly related to their usuabilty as universal dynamical filters when performing general predictive tasks. Second, of course, as in many other areas of science, deeper theory with closed-form expressions for quantities of interests are possible (at least initially) only for sufficiently constrained cases. This is the case

here - the first study of asymptotic behavior of Fisher memory. The linearity and symmetry allowed us to use theory of random matrices and rotate the co-ordinate axis to the natural diagonalised eigen-system.

This study can nevertheless bring some substance to the debate on the "optimal" reservoir construction. While reservoir architecture should reflect the task to be tackled, it is surprising how universal randomized reservoirs can be for a wide variety of tasks. What aspects of dynamic reservoirs are contributing to this universality? If one believes that memory properties can play a role in making dynamical systems good general-purpose dynamical filters, then this study suggests that independent generation of individual reservoir coupling weights can be preferable to (perhaps more sophisticated) dependencies between the weights. It also suggests that, especially for larger reservoirs, self-couplings are much less important than cross-neuron communications. A systematic empirical and theoretical investigations along those lines, possibly making connections to neuroscience, is a matter for future research.

Acknowledgement:
This work was supported by the EPSRC grant "Personalised Medicine Through Learning in the Model Space" (grant number EP/L000296/1).

References

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

[2] 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.

[3] Terry Bossomaier, Lionel Barnett, Michael Harr, and Joseph T. Lizier. *An Introduction to Transfer Entropy: Information Flow in Complex Systems*. Springer Publishing Company, Incorporated, 1st edition, 2016.

[4] Romain Couillet, Gilles Wainrib, Harry Sevi, and Hafiz Tiomoko Ali. The asymptotic performance of linear echo state neural networks. *Journal of Machine Learning Research*, 17(178):1–35, 2016.

[5] J Dambre, David Verstraeten, Benjamin Schrauwen, and Serge Massar. Information processing capacity of dynamical systems. *Scientific reports*, 2:514, 07 2012.

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

[7] Surya Ganguli and Haim Sompolinsky. Short-term memory in neuronal networks through dynamical compressed sensing. In *Advances in neural information processing systems*, pages 667–675, 2010.

[8] Lyudmila Grigoryeva, Julie Henriques, Laurent Larger, and Juan-Pablo Ortega. Nonlinear memory capacity of parallel time-delay reservoir computers in the processing of multidimensional signals. *Neural Computation*, 28(7):1411–1451, 2016.

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

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

[11] Ata Kabán. New bounds on compressive linear least squares regression. In *The 17-th International Conference on Artificial Intelligence and Statistics (AISTATS 2014)*, volume 33, page 448456, 2014.

[12] R. Legenstein and W. Maass. What makes a dynamical system computationally powerful? In S. Haykin, J. C. Principe, T. Sejnowski, and J. McWhirter, editors, *New Directions in Statistical Signal Processing: From Systems to Brains*, pages 127–154. MIT Press, 2007.

[13] Joseph T. Lizier, Mikhail Prokopenko, and Albert Y. Zomaya. Detecting non-trivial computation in complex dynamics. In *Proceedings of the 9th European Conference on Advances in Artificial Life*, ECAL'07, pages 895–904, Berlin, Heidelberg, 2007. Springer-Verlag.

[14] Joseph T. Lizier, Mikhail Prokopenko, and Albert Y. Zomaya. Local measures of information storage in complex distributed computation. *Inf. Sci.*, 208:39–54, November 2012.

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

[16] O. Obst and J. Boedecker. Guided self-organization of input-driven recurrent neural networks. In *In Guided Self-Organization: Inception. Emergence, Complexity and Computation*, pages 319–340. Springer, Berlin, Heidelberg, 2014.

[17] Oliver Obst, Joschka Boedecker, and Minoru Asada. Improving recurrent neural network performance using transfer entropy. In *Proceedings of the 17th International Conference on Neural Information Processing: Models and Applications - Volume Part II*, ICONIP'10, pages 193–200, Berlin, Heidelberg, 2010. Springer-Verlag.

[18] M. Rudelson and R. Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *In Proceedings of the International Congress of Mathematicians. Volume III, Hindustan Book Agency, New Delhi*, pages 1576–1602, 2010.

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

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

[21] 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.