Short Term Memory in Input-Driven Linear Dynamical Systems

Peter Tiňo[a], Ali Rodan[b]

[a]*School of Computer Science, The University of Birmingham*
Birmingham B15 2TT, United Kingdom
E-mail: P.Tino@cs.bham.ac.uk
[b]*King Abdulla II School for Information Technology, University of Jordan*
P.O.Box 11942, Amman, Jordan
E-mail: a.rodan@ju.edu.jo

Abstract

We investigate the relation between two quantitative measures characterizing short term memory in input driven dynamical systems, namely the short term memory capacity (MC) [3] and the Fisher memory curve (FMC) [2]. We show that even though MC and FMC map the memory structure of the system under investigation from two quite different perspectives, for linear input driven dynamical systems they are in fact closely related. In particular, under some assumptions, the two quantities can be interpreted as squared 'Mahalanobis' norms of images of the input vector under the system's dynamics. We also offer a detailed rigorous analysis of the relation between MC and FMC in cases of symmetric and cyclic dynamic couplings.

Keywords: Short term memory capacity, Fisher memory curve, Recurrent neural network, Echo state network, Reservoir Computing

1. Introduction

Input driven dynamical systems have been prominently used as machine learning models in situations when data sets exhibit temporal dependencies, e.g. in time-series prediction [5], speech recognition [8], noise modeling [5], dynamic pattern classification [4], reinforcement learning [1], or language modeling [9]. In an attempt to characterize dynamic properties of such systems, predominantly in the filed of reservoir computation [6], measures have been suggested to quantify how well past information can be represented in the system's internal state. In this contribution we concentrate on two most popular measures of this kind, namely the *short term memory capacity spectrum* MC_k [3] and the *Fisher memory curve* $J(k)$ [2]. These two quantities were derived in different frameworks, mapping the memory structure of the system under investigation from two different perspectives. Perhaps surprisingly, so far their relation has not been closely investigated. In this study we take first steps in this direction and show that under some conditions MC_k and $J(k)$ can be closely related in an interpretable manner.

The paper has the following organization: In section 2 we briefly introduce two memory quantifications considered in this study. Section 3 studies the relation between MC_k and $J(k)$ in the general setting, while in sections 4 and 5 this relation is analyzed in detail for cases of symmetric and cyclic dynamic couplings, respectively. Finally, section 6 summarizes key points of this study.

2. Memory quantifications of linear systems

In this contribution we study linear input driven state space models with N-dimensional state space and univariate inputs and outputs. Such systems can be represented e.g. by linear Echo State Networks (ESN) [6] with N recurrent (reservoir) units. We denote the activations of the input, internal (state), and output units at time step t by $s(t)$, $\mathbf{x}(t)$, and $y(t)$, respectively. The input-to-recurrent and recurrent-to-output unit connections are given by N-dimensional weight vectors $\mathbf{v}$ and $\mathbf{u}$, respectively; connections between the internal units are collected in an $N \times N$ weight matrix W. We assume there are no feedback connections from the output to the dynamic hidden layer (reservoir) and no direct connections from the input to the output. Under these conditions, the state (activation of reservoir units) is updated according to:

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

where $\mathbf{z}(t)$ are zero-mean noise terms. The linear readout is computed as[1]:

$$y(t) = \mathbf{u}^T\mathbf{x}(t). \tag{2}$$

[1] The reservoir activation vector is extended with a fixed element accounting for the bias term.

The output weights $\mathbf{u}$ are typically trained both offline and online by minimizing the Normalized Mean square Error:

$$NMSE = \frac{\langle (y(t) - \tau(t))^2 \rangle}{\langle (\tau(t) - \langle \tau(t) \rangle)^2 \rangle}, \tag{3}$$

where $\tau(t)$ is the desired output (target) at time t and $\langle \cdot \rangle$ denotes the empirical mean.

In ESN, the elements of W and $\mathbf{v}$ are fixed prior to training with random values drawn from a uniform distribution over a (typically) symmetric interval. The reservoir connection matrix W is typically scaled as $W \leftarrow \omega W / |\lambda_{max}|$, where $|\lambda_{max}|$ is the spectral radius of W and $0 < \omega < 1$ is a scaling parameter [6].

Even though memory quantifications studied in this paper have been developed in the context of reservoir computation (in particular, ESN), their theory is general and can be used for any smooth input-driven dynamical system with univariate input. We will therefore develop the theory in the context of eq. (1) and mention ESN for historical reasons only.

2.1. Short Term Memory Capacity (MC)

In [3] Jaeger quantified the ability of recurrent network architectures to represent past events through a measure correlating the past events in a (typically i.i.d.) input stream with the network output. In particular, the network (1) without dynamic noise ($\mathbf{z}(t) = \mathbf{0}$) is driven by a univariate stationary input signal $s(t)$. For a given delay k, we consider the network with optimal parameters for the task of outputting $s(t-k)$, after seeing the input stream $...s(t-1)s(t)$ up to time t. The goodness of fit is measured in terms of the squared correlation coefficient between the desired output $\tau(t) = s(t-k)$ and the observed network output $y(t)$:

$$MC_k = \frac{Cov^2(s(t-k), y(t))}{Var(s(t))\ Var(y(t))}, \tag{4}$$

where Cov and Var denote the covariance and variance operators, respectively. Hence, this quantity quantifies how much information about the value of the input signal from k time steps back, $s(t-k)$, can be inferred (through linear readout (2)) from the current state $\mathbf{x}(t)$. The memory is tested to the extreme as the inputs are mutually independent and so no information about the other inputs $s(t)$, ..., s(t-k+1), $s(t-k-1), \ldots$ could be useful for the task of recalling $s(t-k)$.

The short term memory (STM) capacity is then given by [3]

$$MC = \sum_{k=1}^{\infty} MC_k. \tag{5}$$

Jaeger [3] proved that for *any* recurrent neural network with N recurrent neurons, under the assumption of i.i.d. input stream, the STM capacity cannot exceed N.

2.2. Fisher Memory Curve (FMC)

Memory capacity MC represents one particular way of quantifying the amount of information that can be preserved in the state space about the past inputs. In [2] Ganguli, Huh and Sompolinsky proposed a different quantification of 'memory' for linear input driven dynamical systems corrupted by a Gaussian state noise. In particular, it is assumed that the dynamic noise $\mathbf{z}(t)$ is a memoryless process of i.i.d. zero mean Gaussian variables with covariance ϵI, $\epsilon > 0$, where I is the identity matrix. Under such dynamic noise, given an input driving stream $s(..t) = \ldots\ s(t-2)\ s(t-1)\ s(t)$, the input-conditional state distribution $p(\mathbf{x}(t)|s(..t))$ is a Gaussian with covariance [2]

$$C = \epsilon \sum_{\ell=0}^{\infty} W^{\ell} (W^T)^{\ell}. \tag{6}$$

The Fisher memory matrix quantifies sensitivity 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),

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

and its diagonal elements $J(k) = F_{k,k}(s(..t))$ quantify the information that the state distribution retains about a change (e.g. a pulse) entering the network k time steps in the past. The collection of terms $\{J(k)\}_{k=0}^{\infty}$ was termed Fisher memory curve (FMC) and evaluated to [2]

$$J(k) = \mathbf{v}^T (W^T)^k C^{-1} W^k \mathbf{v}. \tag{7}$$

Note that in the Fisher memory framework, the input sequences are considered "parameter settings" in the sense that each input sequence parameterizes the mean of the state distribution characterizing the state of the system after processing that input sequence. Expected sensitivities of the state distributions to small perturbations in input sequences ('parameters'), quantified by Fisher information, are then used as indicators of how much small past changes in the input sequences get reflected in changes in the resulting state distributions.

We stress that, unlike in the case of short term memory capacity, FMC does not depend in any way on the input driving stream. This is because the system is linear and the dynamic noise is i.i.d. with fixed variance. The mean of the state-conditional Gaussian distribution is linear in the input sequence, making the Fisher information independent of input sequence perturbations. The memory capacity can be increased by adding temporal dependencies in the input stream $s(..t)$, but to get to the very essence of the system's memory, the standard practice is to impose no dependencies in $s(..t)$. On the other hand, by construction of FMC, the terms $J(k)$ do not depend on memory aspects of the input driving source and Fisher memory cannot be boosted by any additional information about the other inputs presented so far.

3. Short term memory capacity and Fisher memory curve – general case

In this section we will investigate the relation between short term memory capacity MC_k and Fisher memory curve $J(k)$, without assuming any specific structure of the state space model in terms of the dynamic coupling (parameter matrix) W.

We first briefly introduce some necessary notation. Denote the image of the input weight vector $\mathbf{v}$ through k-fold application of the dynamic operator W by $\mathbf{v}^{(k)}$, i.e. $\mathbf{v}^{(k)} = W^k \, \mathbf{v}$. Provided the matrix

$$A = \frac{1}{\epsilon} C - G, \tag{8}$$

where

$$G = \sum_{\ell=0}^{\infty} \mathbf{v}^{(\ell)} \; (\mathbf{v}^{(\ell)})^T, \tag{9}$$

is invertible, we define

$$D = G \; (A^{-1} + G^{-1}) \; G. \tag{10}$$

Furthermore, for any positive definite matrix $B \in \mathbb{R}^{n \times n}$ we denote the induced norm on $\mathbb{R}^n$ by $\| \cdot \|_B$, i.e. for any $\mathbf{v} \in \mathbb{R}^n$, $\|\mathbf{v}\|_B^2 = \mathbf{v}^T B \mathbf{v}$. We are now ready to formulate the main result.

Theorem 1: *Let MC_k be the k-th memory capacity term (4) of the system (1) with no dynamic noise, under a zero-mean i.i.d. input driving source. Let $J(k)$ be the k-th term of the Fisher memory curve (7) of the system (1) with i.i.d. dynamic noise of variance ϵ. If D is positive definite, then*

$$MC_k = \epsilon \ J(k) + \|\mathbf{v}^{(k)}\|_{D^{-1}}^2 \tag{11}$$

and $MC_k > \epsilon \ J(k)$, for all $k > 0$.

Proof: Given an i.i.d. zero-mean real-valued input stream $s(..t) = ... \ s(t-2) \ s(t-1) \ s(t)$ of variance σ^2 emitted by a source P, the state at time t of the system (1) (under no dynamic noise ($\epsilon = 0$)) is

$$\mathbf{x}(t) = \sum_{\ell=0}^{\infty} s(t-\ell) \ W^\ell \ \mathbf{v} = \sum_{\ell=0}^{\infty} s(t-\ell) \ \mathbf{v}^{(\ell)}.$$

For the task of recalling the input from k time steps back, the optimal least-squares readout vector $\mathbf{u}$ is given by

$$\mathbf{u} = R^{-1} \ \mathbf{p}^{(k)}, \tag{12}$$

where

$$R = E_{P(s(..t))}[\mathbf{x}(t) \ \mathbf{x}^T(t)] = \sigma^2 \ G$$

is the covariance matrix of reservoir activations and

$$\mathbf{p}^{(k)} = E_{P(s(..t))}[s(t-k) \ \mathbf{x}(t)] = \sigma^2 \ \mathbf{v}^{(k)}.$$

Provided R is full rank, the optimal readout vector $\mathbf{u}^{(k)}$ for delay k reads

$$\mathbf{u}^{(k)} = G^{-1} \ \mathbf{v}^{(k)}. \tag{13}$$

The optimal 'recall' output at time t is then $y(t) = \mathbf{x}^T(t) \ \mathbf{u}^{(k)}$, yielding

$$Cov(s(t-k), y(t)) \ = \ \sigma^2 \ (\mathbf{v}^{(k)})^T \ G^{-1} \ \mathbf{v}^{(k)}. \tag{14}$$

Since for the optimal recall output $Cov(s(t-k), y(t)) = Var(y(t))$ [3, 7], we have

$$MC_k = (\mathbf{v}^{(k)})^T \ G^{-1} \ \mathbf{v}^{(k)}. \tag{15}$$

Rewriting (7) as

$$J(k) = (\mathbf{v}^{(k)})^T C^{-1} \mathbf{v}^{(k)},$$

it is noticeable that the Fisher memory curve $J(k)$ and memory capacity terms MC_k have the same form. The matrix $G = \sum_{\ell=0}^{\infty} \mathbf{v}^{(\ell)} \ (\mathbf{v}^{(\ell)})^T$ can be considered a scaled 'covariance' matrix of the iterated images of $\mathbf{v}$ under the state space mapping. Then MC_k is the squared 'Mahalanobis norm' of $\mathbf{v}^{(k)}$ under the covariance structure G,

$$MC_k \ = \ \|\mathbf{v}^{(k)}\|^2_{G^{-1}}. \tag{16}$$

Analogously, $J(k)$ is the squared 'Mahalanobis norm' of $\mathbf{v}^{(k)}$ under the covariance C of the state distribution $p(\mathbf{x}(t)|s(..t))$ induced by the dynamic noise $\mathbf{z}(t)$,

$$J(k) \ = \ \|\mathbf{v}^{(k)}\|^2_{C^{-1}}. \tag{17}$$

Denote the rank-1 matrix $\mathbf{v}\mathbf{v}^T$ by Q. From (6), (8) and (9) we have

$$A = \sum_{\ell=0}^{\infty} W^{\ell} \ (I - Q) \ (W^T)^{\ell}.$$

Furthermore (from (8)),

$$\epsilon C^{-1} = (A + G)^{-1}$$

and, provided A is invertible (and $(A^{-1}+G^{-1})$ is invertible as well), by matrix inversion lemma, we obtain

$$\epsilon C^{-1} = G^{-1} - G^{-1} \ (A^{-1} + G^{-1})^{-1} \ G^{-1}.$$

This leads to

$$\begin{aligned} J(k) \ &= \ (\mathbf{v}^{(k)})^T \ C^{-1} \ \mathbf{v}^{(k)} \\ &= \ \frac{1}{\epsilon} MC_k - \frac{1}{\epsilon} (\mathbf{v}^{(k)})^T \ D^{-1} \ \mathbf{v}^{(k)}. \end{aligned}$$

Since G and A are symmetric matrices, so are their inverses and hence D is also a symmetric matrix. Provided D is positive definite, it can be considered (inverse of a) metric tensor and

$$MC_k = \epsilon \ J(k) + \|\mathbf{v}^{(k)}\|^2_{D^{-1}}.$$

Obviously, in such a case, $MC_k > \epsilon\ J(k)$ for all $k > 0$. □

We do not elaborate here on exactly under what conditions will the matrix D be positive definite. Determining such conditions can be quite tricky and would require a separate study. Nevertheless, we generated many instances of the system (1) by random sampling of W and $\mathbf{v}$ and re-normalizing W as in ESN (see section 2). For at least half of the generated model instances the matrix D was indeed positive definite, and for those input-driven dynamical models the relationship between the memory quantifications stated in Theorem 1 applies. As an illustrative example, we present in figure 1 the norm $\|\mathbf{v}^{(k)}\|^2_{D^{-1}}$, $k = 1, 2, ..., 20$, that makes up for the difference between the terms of memory capacity (MC_k) and Fisher memory ($J(k)$). Randomly constructed 12-dimensional system (1) was used, coupling weights were randomly generated from a uniform distribution over an interval symmetric around zero and then W was normalized to spectral radius 0.99. Input weights were generated from uniform distribution over $[-0.5, 0.5]$ and the input vector was renormalized to L_2 norm 1. The norm eventually decays due to increasingly contractive character of the linear operator W^k transforming $\mathbf{v}$ into $\mathbf{v}^{(k)}$.

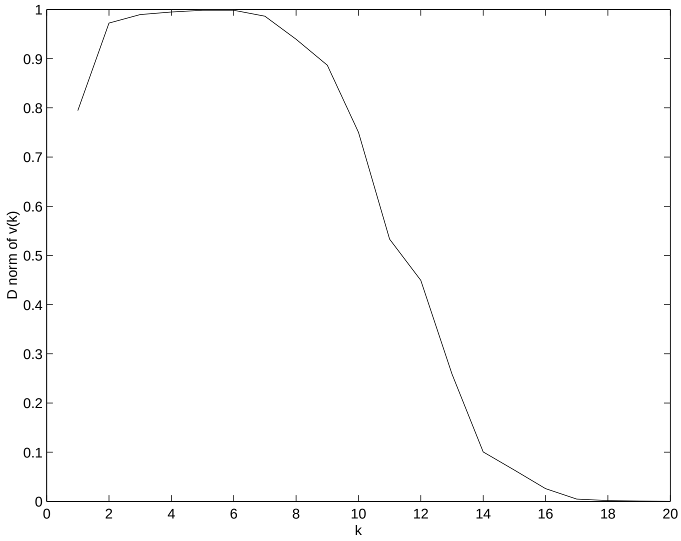

Figure 1: Norm $\|\mathbf{v}^{(k)}\|^2_{D^{-1}}$, $k = 1, 2, ..., 20$, that makes up for the difference between the memory capacity MC_k and Fisher memory $J(k)$ (see Theorem 1). Randomly constructed 12-dimensional system (1) was used.

4. Linear systems with symmetric *W*

In this section we will derive the relationship between the two memory measures, Fisher Memory Curve and Memory Capacity, for a class of input driven systems characterized by symmetric coupling matrices W. We will assume that W is full rank - treating lower rank systems would be analogous. We will further assume that the system is not trivial, e.g. $W \neq I$.

4.1. Fisher Memory Curve

For symmetric W, covariance matrix of the state distribution $p(\mathbf{x}(t)|s(..t))$ (6) can be written as

$$\begin{aligned} C &= \epsilon \sum_{\ell=0}^{\infty} W^{\ell} (W^T)^{\ell} \\ &= \epsilon \sum_{\ell=0}^{\infty} (W^2)^{\ell} \\ &= \epsilon (I - W^2)^{-1} \end{aligned} \tag{18}$$

and $C^{-1} = \epsilon^{-1} \; (I - W^2)$.

Consider diagonalization of W,

$$W = U \Lambda U^T, \quad \Lambda = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_N), \tag{19}$$

U orthonormal, so that $W^k = U \Lambda^k U^T$.

We can now evaluate elements of the Fisher memory curve (see (7)):

$$\begin{aligned} J(k) &= \frac{1}{\epsilon} \mathbf{v}^T U \Lambda^k U^T (I - W^2) U \Lambda^k U^T \mathbf{v} \\ &= \frac{1}{\epsilon} \mathbf{v}^T U \Lambda^k [U^T - \Lambda^2 U^T] U \Lambda^k U^T \mathbf{v} \\ &= \frac{1}{\epsilon} \mathbf{v}^T U \Lambda^k [I - \Lambda^2] \Lambda^k U^T \mathbf{v} \\ &= \frac{1}{\epsilon} \tilde{\mathbf{v}}^T \Lambda^{2k} [I - \Lambda^2] \tilde{\mathbf{v}}, \end{aligned} \tag{20}$$

where $\tilde{\mathbf{v}} = U^T \mathbf{v}$ is the expression of $\mathbf{v}$ in the eigenbasis U. If the orthonormal basis (columns of U) is $\mu_1, \mu_1, ..., \mu_N$, then $\mathbf{v} = \sum_{i=1}^{N} \tilde{v}_i \mu_i$.

Denote by Λ_0 the diagonal matrix $\Lambda^{2k}[I - \Lambda^2]$, e.g.

$$\Lambda_0 = \mathrm{diag}(\lambda_1^{2k}(1 - \lambda_1^2), ..., \lambda_N^{2k}(1 - \lambda_N^2)), \tag{21}$$

and denote by $\mathbf{e}_i$ 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.
Then, from (20-21),

$$\begin{aligned} J(k) &= \frac{1}{\epsilon}\left[\sum_{i=1}^{N} \tilde{v}_i \mathbf{e}_i^T\right] \Lambda_0 \left[\sum_{j=1}^{N} \tilde{v}_j \mathbf{e}_j\right] \\ &= \frac{1}{\epsilon}\sum_{i,j=1}^{N} \tilde{v}_i \tilde{v}_j \lambda_j^{2k}(1-\lambda_j^2)\mathbf{e}_i^T \mathbf{e}_j \\ &= \frac{1}{\epsilon}\sum_{i=1}^{N} \tilde{v}_i^2 \lambda_i^{2k}(1-\lambda_i^2). \end{aligned} \tag{22}$$

4.2. Memory Capacity

Using the diagonalization (19) of W, we calculate

$$\begin{aligned} G &= \sum_{\ell=0}^{\infty} W^{\ell} \mathbf{v}\mathbf{v}^T (W)^{\ell} \\ &= \sum_{\ell=0}^{\infty} U\Lambda^{\ell} U^T \mathbf{v}\mathbf{v}^T U \Lambda^{\ell} U^T \\ &= \sum_{\ell=0}^{\infty} U\Lambda^{\ell} \tilde{\mathbf{v}}\tilde{\mathbf{v}}^T \Lambda^{\ell} U^T \\ &= U\left[\sum_{\ell=0}^{\infty} \Lambda^{\ell}\left[\sum_{i,j=1}^{N} \tilde{v}_i \tilde{v}_j \mathbf{e}_i \mathbf{e}_j^T\right]\Lambda^{\ell}\right] U^T. \end{aligned} \tag{23}$$

Denote by $T^{(i,j)}$ the $N \times N$ matrix $\mathbf{e}_i \mathbf{e}_j^T$ that has all 0's, except for 1 at position (i,j). Then,

$$G = U\left[\sum_{\ell=0}^{\infty}\sum_{i,j=1}^{N} \tilde{v}_i \tilde{v}_j \Lambda^{\ell} T^{(i,j)} \Lambda^{\ell}\right] U^T. \tag{24}$$

Note that $\Lambda^{\ell} T^{(i,j)} \Lambda^{\ell}$ is a $\mathbf{0}$ matrix if $i \neq j$. Hence,

$$G = U\left[\sum_{\ell=0}^{\infty}\sum_{i=1}^{N} \tilde{v}_i^2 \; \lambda_i^{2\ell} \; T^{(i,i)}\right] U^T$$

$$
\begin{aligned}
&= U \left[\sum_{i=1}^{N} \tilde{v}_i^2 \; T^{(i,i)} \; \sum_{\ell=0}^{\infty} \lambda_i^{2\ell} \right] U^T \\
&= U \left[\sum_{i=1}^{N} \tilde{v}_i^2 \; T^{(i,i)} \; \frac{1}{1-\lambda_i^2} \right] U^T \\
&= U \; \mathrm{diag} \left(\tilde{v}_1^2 \frac{1}{1-\lambda_1^2}, ..., \tilde{v}_N^2 \frac{1}{1-\lambda_N^2} \right) \; U^T, \qquad (25)
\end{aligned}
$$

and so

$$G^{-1} = U \; \mathrm{diag}(\tilde{v}_1^{-2}(1-\lambda_1^2), ..., \tilde{v}_N^{-2}(1-\lambda_N^2)) \; U^T. \qquad (26)$$

Denote $\mathrm{diag}(\tilde{v}_1^{-2}, ..., \tilde{v}_1^{-2})$ by $\tilde{V}$. Then,

$$G^{-1} = U \; \tilde{V} \; [I - \Lambda^2] \; U^T \qquad (27)$$

and

$$
\begin{aligned}
MC_k &= \mathbf{v}^T W^k G^{-1} W^k \mathbf{v} \\
&= \mathbf{v}^T U \Lambda^k \tilde{V} [I - \Lambda^2] \Lambda^k U^T \mathbf{v} \\
&= \tilde{\mathbf{v}}^T \Lambda^k \tilde{V} [I - \Lambda^2] \Lambda^k \tilde{\mathbf{v}}. \qquad (28)
\end{aligned}
$$

Since

$$\Lambda^k \tilde{V} [I - \Lambda^2] \Lambda^k \;\; = \;\; \mathrm{diag}(\lambda_1^{2k}(1-\lambda_1^2)\tilde{v}_1^{-2}, ..., \lambda_N^{2k}(1-\lambda_N^2)\tilde{v}_N^{-2}), \qquad (29)$$

we have

$$
\begin{aligned}
MC_k &= \sum_{i=1}^{N} \tilde{v}_i^2 \lambda_i^{2k} (1-\lambda_i^2) \tilde{v}_i^{-2} \\
&= \sum_{i=1}^{N} \lambda_i^{2k} (1-\lambda_i^2) \qquad (30)
\end{aligned}
$$

4.3. Short term memory capacity and Fisher memory curve – symmetric W

In the above sections we have proved the following theorem that summarizes the relation between short term memory capacity and Fisher memory curve for linear systems with non-trivial symmetric coupling.

Theorem 2: *Let MC_k be the k-th memory capacity term (4) of the system (1) with no dynamic noise, under a zero-mean i.i.d. input driving source.*

Let $J(k)$ be the k-th term of the Fisher memory curve (7) of the system (1) with i.i.d. dynamic noise of variance ϵ. If W is a symmetric regular matrix W, $W \neq I$, then

$$J(k) = \frac{1}{\epsilon}\sum_{i=1}^{N} \tilde{v}_i^2 \lambda_i^{2k}(1-\lambda_i^2), \tag{31}$$

and

$$MC_k = \sum_{i=1}^{N} \lambda_i^{2k}(1-\lambda_i^2). \tag{32}$$

In particular, we have the following Corollary:

Corollary 3: *Under the assumptions of Theorem 2, if $\tilde{v}_i = \sqrt{\epsilon}$, $i = 1, 2, ..., N$, then*

$$J(k) = MC_k. \tag{33}$$

One can elaborate on theorem 2 further and assume that the input vector $\mathbf{v}$ is collinear with the sum of the eigenvectors of W,

$$\mathbf{v} = \alpha \sum_{i=1}^{N} \mu_i, \tag{34}$$

where $\alpha > 0$. Then

$$\begin{aligned}\tilde{\mathbf{v}} &= U^T\mathbf{v} = \alpha\sum_{i=1}^{N} U^T \mu_i \\ &= \alpha\ \mathbf{1}_N, \end{aligned}\tag{35}$$

where $\mathbf{1}_N$ is the N-dimensional vector of 1's. Hence,

$$J(k) = \frac{\alpha^2}{\epsilon}\sum_{i=1}^{N} \lambda_i^{2k}(1-\lambda_i^2), \tag{36}$$

and so,

$$J(k) \;=\; \frac{\alpha^2}{\epsilon} MC_k. \tag{37}$$

Corollary 4: *Under the assumptions of Theorem 2, if the input vector* $\mathbf{v}$ *is collinear with the sum of the eigenvectors of* W*,* $\mathbf{v} = \alpha \sum_{i=1}^{N} \mu_i$*,* $\alpha > 0$*, then*

$$J(k) = \frac{\alpha^2}{\epsilon} MC_k. \tag{38}$$

5. Linear systems with scaled orthonormal *W*

Finally, we turn our attention to linear systems with scaled orthonormal W, i.e. there is an orthonormal matrix H such that $W = rH$, for some $r > 0$. Then, $WW^T = r^2 I$. An example of such a system was studied in detail in [7] under the name *Simple Cycle Reservoir* (SCR). In SCR the couplings have the cyclic form

$$W = r \cdot [\mathbf{e}_2, \mathbf{e}_3, ..., \mathbf{e}_N, \mathbf{e}_1]$$

and $0 < r < 1$.

The memory capacity spectrum of SCR was evaluated to [7]

$$MC_{qN+j} \;=\; (1 - r^{2N}) r^{2Nq}, \quad q = 0, 1, 2, ..., \quad j = 1, 2, ..., N, \tag{39}$$

i.e. the MC terms decay in a piece-wise constant manner.

We now evaluate the Fisher memory curve of such systems. Covariance of the state distribution reads

$$\begin{aligned} C &= \epsilon \sum_{\ell=0}^{\infty} W^{\ell} (W^T)^{\ell} \\ &= \epsilon \sum_{\ell=0}^{\infty} W...W \; W \; W^T \; W^T...W^T \\ &= \epsilon \left[\sum_{\ell=0}^{\infty} r^{2\ell} \right] I \\ &= \epsilon \frac{1}{1 - r^2} I, \end{aligned} \tag{40}$$

leading to

$$C^{-1} \quad = \quad \frac{1}{\epsilon}(1-r^2)I \tag{41}$$

We are now ready to write the Fisher memory terms:

$$\begin{aligned} J(k) &= \mathbf{v}^T(W^T)^k C^{-1} W^k \mathbf{v} \\ &= \frac{1}{\epsilon}(1-r^2)\mathbf{v}^T W^T ... W^T \ W^T \ W \ W ... W \ \mathbf{v} \\ &= \frac{1}{\epsilon}(1-r^2)\mathbf{v}^T \left(r^{2k}\right) I \ \mathbf{v} \\ &= \frac{1}{\epsilon}(1-r^2) r^{2k} \mathbf{v}^T \mathbf{v} \\ &= \frac{1}{\epsilon} r^{2k}(1-r^2)\|\mathbf{v}\|_E^2, \end{aligned} \tag{42}$$

where $\|\cdot\|_E$ denotes the Euclidean norm. Hence,

$$J(qN+j) \quad = \quad \frac{1}{\epsilon}(1-r^2) r^{2qN+2j}\|\mathbf{v}\|_E^2. \tag{43}$$

For $q = 0, 1, 2, ...$ and $j = 1, 2, ..., N$, the ratio between the Fisher memory and memory capacity terms evaluates to

$$\begin{aligned} \frac{J(qN+j)}{MC_{qN+j}} &= \frac{\|\mathbf{v}\|_E^2}{\epsilon} \, \frac{(1-r^2)}{(1-r^N)} \, r^{2j} \\ &= \kappa_N(r) \ r^{2j} \end{aligned} \tag{44}$$

where the constant term

$$\kappa_N(r) \quad = \quad \frac{\|\mathbf{v}\|_E^2}{\epsilon} \, \frac{(1-r^2)}{(1-r^N)} \tag{45}$$

expresses the “core ratio” between the two memory quantifications, that is further scaled by r^{2j} in each N-block of delays k. For large networks, $r^N \approx 0$ and so $\kappa_N(r) = \epsilon^{-1} \ (1-r^2) \ \|\mathbf{v}\|_E^2$.

6. Conclusions

We investigated the relation between two quantitative measures suggested in the literature to characterize short term memory in input driven dynamical systems, namely the short term memory capacity spectrum MC_k [3] and the Fisher memory curve $J(k)$ [2], for time lags $k \geq 0$. The two memory measures have been formulated in two different frameworks. $J(k)$ is independent of the input driving stream $s(..t)$ and measures the 'inherent' memory capabilities of such systems by measuring the sensitivity of the state distribution $p(\mathbf{x}(t)|s(..t))$ induced by the dynamic noise with respect to perturbations in $s(..t)$, k time steps back. On the other hand MC_k quantifies how well the past inputs $s(t-k)$ can be reconstructed by linearly projecting the state vector $\mathbf{x}(t)$.

We have shown that even though MC_k and $J(k)$ map the memory structure of the system under investigation from two quite different perspectives, they are closely related. In particular, under some assumptions, the two quantities can be interpreted as squared 'Mahalanobis' norms of images of the input vector under the system's dynamics. Furthermore, we have shed more light on the closely related nature of the two memory measures by rigorous analysis of the relation between MC_k and $J(k)$ in cases of symmetric and cyclic dynamic couplings in linear input driven systems.

Acknowledgments

Peter Tiňo was supported by a BBSRC grant (no. BB/H012508/1).

References

[1] K. Bush and C. Anderson. Modeling reward functions for incomplete state representations via echo state networks. In *Proceedings of the International Joint Conference on Neural Networks, Montreal, Quebec*, July 2005.

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

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

[4] H. Jaeger. A tutorial on training recurrent neural networks, covering bppt, rtrl, ekf and the "echo state network" approach. Technical report

gmd report 159, German National Research Center for Information Technology, 2002.

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

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

[7] A. Rodan and P. Tino. Minimum complexity echo state network. *IEEE Transactions on Neural Networks*, 22(1):131–144, 2011.

[8] M.D. Skowronski and J.G. Harris. Minimum mean squared error time series classification using an echo state network prediction model. In *IEEE International Symposium on Circuits and Systems, Island of Kos, Greece, pp. 3153-3156*, 2006.

[9] 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.