ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Short Term Memory Quantifications in Input-Driven Linear Dynamical Systems

Peter Tiňo and Ali Rodan

School of Computer Science, The University of Birmingham
Birmingham B15 2TT, United Kingdom
E-mail: {P.Tino, a.a.rodan}@cs.bham.ac.uk

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) [2] and the Fisher memory curve (FMC) [1]. We show that under some assumptions, the two quantities can be interpreted as squared 'Mahalanobis' norms of images of the input vector under the system's dynamics and that even though MC and FMC map the memory structure of the system from two quite different perspectives, they can be linked by a close relation.

1 Introduction

Input driven dynamical systems play an important role as machine learning models when data sets exhibit temporal dependencies, e.g. in prediction or control. In an attempt to characterize dynamic properties of such systems, measures have been suggested to quantify how well past information can be represented in the system's internal state. In this contribution we investigate two such well known measures, namely the short term memory capacity spectrum MC_k [2] and the Fisher memory curve $J(k)$ [1]. The two quantities map the memory structure of the system under investigation from two quite different perspectives. So far their relation has not been closely investigated. In this paper we take the first step to bridge this gap and show that under some conditions MC_k and $J(k)$ can be related in an interpretable manner.

2 Background

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) [3] with N recurrent (reservoir) units. The activations of the input, internal (state), and output units at time step t are denoted 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 reservoir and no direct connections from the input to the output. Under

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

these conditions, the reservoir units are 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]:

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

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

$$NMSE = \frac{\langle \|y(t) - \tau(t)\|_2^2 \rangle}{\langle \|\tau(t) - \langle \tau(t) \rangle\|_2^2 \rangle}, \tag{3}$$

where $y(t)$ is the readout output, $\tau(t)$ is the desired output (target), $\|.\|_2$ denotes the Euclidean norm 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 $W \leftarrow \alpha W/|\lambda_{max}|$, where $|\lambda_{max}|$ is the spectral radius of W and $0 < \alpha < 1$ is a scaling parameter [3].

Short Term Memory Capacity (MC): In [2] 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 ($\mathbf{z}(t) = 0$) 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. The short term memory (STM) capacity is then given by [2] $MC = \sum_{k=1}^{\infty} MC_k$. Jaeger [2] 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.

Fisher Memory Curve (FMC): Memory capacity MC represents one way of quantifying the amount of information that can be preserved in the reservoir about the past inputs. In [1] Ganguli, Huh and Sompolinsky proposed a

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

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

different quantification of memory capacity for linear reservoirs 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 (I is the identity matrix). Then, given an input driving stream $s(..t) = ... \; s(t-2) \; s(t-1) \; s(t)$, the dynamic noise induces a state distribution $p(\mathbf{x}(t)|s(..t))$, which is a Gaussian with covariance [1]

$$C = \epsilon \sum_{\ell=0}^{\infty} W^{\ell}(W^T)^{\ell}. \tag{5}$$

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 recurrent network are 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 $\mathbf{x}(t)$ retain 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 [1]

$$J(k) = \mathbf{v}^T(W^T)^k C^{-1} W^k \mathbf{v}. \tag{6}$$

Note that, unlike the short term memory capacity, it turns out that FMC does not depend on the input driving stream.

3 Relation between short term memory capacity and Fisher memory curve

We first briefly introduce some necessary notation. Denote the image of the input weight vector $\mathbf{v}$ through k-fold application of the reservoir operator W by $\mathbf{v}^{(k)}$, i.e. $\mathbf{v}^{(k)} = W^k \; \mathbf{v}$. Define $A = \frac{1}{\epsilon}C - G$, where

$$G = \sum_{\ell=0}^{\infty} \mathbf{v}^{(\ell)} \; (\mathbf{v}^{(\ell)})^T. \tag{7}$$

Provided A is invertible, denote $G \; (A^{-1} + G^{-1}) \; G$ by D. 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: *Let MC_k be the k-th memory capacity term (4) of network (1) with no dynamic noise, under a zero-mean i.i.d. input driving source. Let $J(k)$ be*

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

the k-th term of the Fisher memory curve (6) of network (1) with i.i.d. dynamic noise of variance ϵ. If D is positive definite, then

$$MC_k = \epsilon \; J(k) + \|\mathbf{v}^{(k)}\|^2_{D^{-1}} \tag{8}$$

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 linear reservoir (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

$$\mathbf{u} = R^{-1} \; \mathbf{p}^{(k)}, \tag{9}$$

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

$$\mathbf{u}^{(k)} = G^{-1} \; \mathbf{v}^{(k)}. \tag{10}$$

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

$$Cov(s(t-k), y(t)) \quad = \quad \sigma^2 \; (\mathbf{v}^{(k)})^T \; G^{-1} \; \mathbf{v}^{(k)}. \tag{11}$$

Since for the optimal recall output $Cov(s(t-k), y(t)) = Var(y(t))$ [2, 4], we have

$$MC_k = (\mathbf{v}^{(k)})^T \; G^{-1} \; \mathbf{v}^{(k)}. \tag{12}$$

The Fisher memory curve and memory capacity terms (6) and (12), respectively 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 reservoir mapping. Then MC_k is the squared 'Mahalanobis norm' of $\mathbf{v}^{(k)}$ under the covariance structure G,

$$MC_k \quad = \quad \|\mathbf{v}^{(k)}\|^2_{G^{-1}}. \tag{13}$$

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

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 $\mathbf{z}(t)$,

$$\begin{aligned} J(k) &= (\mathbf{v}^{(k)})^T \; C^{-1} \; \mathbf{v}^{(k)} \\ &= \|\mathbf{v}^{(k)}\|^2_{C^{-1}}. \end{aligned} \tag{14}$$

Denote the rank-1 matrix $\mathbf{v}\mathbf{v}^T$ by Q. Then by (5), $\frac{1}{\epsilon}C = A + G$, where

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

It follows that $\epsilon C^{-1} = (A+G)^{-1}$ and, provided A is invertible (and $(A^{-1}+G^{-1})$ is invertible as well), by matrix inversion lemma,

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

We have

$$\begin{aligned} J(k) &= (\mathbf{v}^{(k)})^T \; C^{-1} \; \mathbf{v}^{(k)} \\ &= \frac{1}{\epsilon} M C_k - \frac{1}{\epsilon} (\mathbf{v}^{(k)})^T \; D^{-1} \; \mathbf{v}^{(k)}, \end{aligned}$$

where

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

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

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

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

From (8) we have: $\sum_{k=0}^{\infty} MC_k = \epsilon \; \sum_{k=0}^{\infty} J(k) + \sum_{k=0}^{\infty} \|\mathbf{v}^{(k)}\|^2_{D^{-1}}$. If the input weight vector $\mathbf{v}$ is a unit vector ($\|\mathbf{v}\|_2 = 1$) and the reservoir matrix W is normal (i.e. has orthogonal eigenvector basis), we have $\sum_{k=0}^{\infty} J(k) = 1$ [1]. In such cases $\sum_{k=0}^{\infty} MC_k = N$, implying

$$\sum_{k=0}^{\infty} \|\mathbf{v}^{(k)}\|^2_{D^{-1}} = N - \epsilon. \tag{15}$$

As an example of metric structures underlying the norms in (8), (13) and (14), we show in figure 1 covariance structure of C ($\epsilon = 1$), G and D corresponding to a 15-node linear reservoir. The covariances were projected onto the two-dimensional space spanned by the 1st and 14th eigenvectors of C (rank determined by decreasing eigenvalues). Reservoir weights were randomly generated from a uniform distribution over an interval symmetric around zero and then W was normalized to spectral radius 0.995. Input weights were generated from uniform distribution over $[-0.5, 0.5]$.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

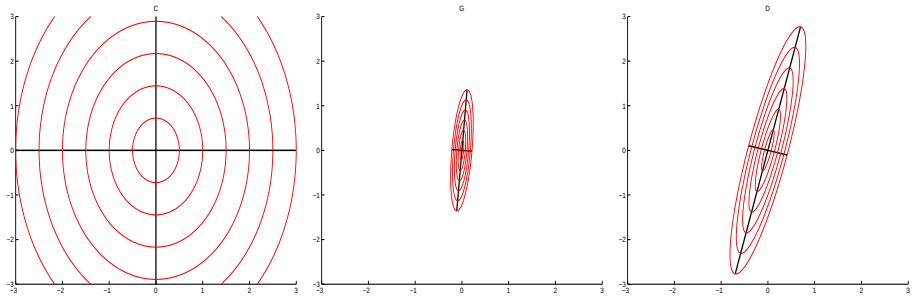

Fig. 1: Covariance structure of C (left), G (middle) and D (right) for a 15-node linear reservoir projected onto the 1st and 14th eigenvectors of C. Shown are iso-lines corresponding to 0.5, 1, 1.5, ..., 3 standard deviations.

4 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 and the Fisher memory curve $J(k)$, for time lags $k \geq 0$. $J(k)$ 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 under some assumptions, the two quantities can be interpreted as squared 'Mahalanobis' norms of images of the input vector under the system's dynamics and that $MC_k > \epsilon\ J(k)$, for all $k > 0$. Even though MC_k and $J(k)$ map the memory structure of the system under investigation from two quite different perspectives, they can be closely related.

References

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

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

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

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