Journal of Machine Learning Research 25 (2024) 1-38 Submitted 5/23; Revised 7/24; Published 8/24

Memory of Recurrent Networks: Do We Compute It Right?

Giovanni Ballarin GIOVANNI.BALLARIN@GESS.UNI-MANNHEIM.DE
Department of Economics
University of Mannheim
Germany

Lyudmila Grigoryeva LYUDMILA.GRIGORYEVA@UNISG.CH
Faculty of Mathematics and Statistics
Universität Sankt Gallen
Switzerland
Department of Statistics (Honorary Assoc. Prof.)
University of Warwick
United Kingdom

Juan-Pablo Ortega JUAN-PABLO.ORTEGA@NTU.EDU.SG
Division of Mathematical Sciences
School of Physical and Mathematical Sciences
Nanyang Technological University
Singapore

Editor: Sayan Mukherjee

Abstract

Numerical evaluations of the memory capacity (MC) of recurrent neural networks reported in the literature often contradict well-established theoretical bounds. In this paper, we study the case of linear echo state networks, for which the total memory capacity has been proven to be equal to the rank of the corresponding Kalman controllability matrix. We shed light on various reasons for the inaccurate numerical estimations of the memory, and we show that these issues, often overlooked in the recent literature, are of an exclusively numerical nature. More explicitly, we prove that when the Krylov structure of the linear MC is ignored, a gap between the theoretical MC and its empirical counterpart is introduced. As a solution, we develop robust numerical approaches by exploiting a result of MC neutrality with respect to the input mask matrix. Simulations show that the memory curves that are recovered using the proposed methods fully agree with the theory.

Keywords: reservoir computing, linear recurrent neural networks, echo state networks, memory capacity, Krylov iterations

1. Introduction

Recurrent Neural Networks (RNNs) are among the most widely used machine learning tools for sequential data processing (Sutskever et al., 2014). Despite the rising popularity of transformer deep neural architectures (Vaswani et al., 2017; Galimberti et al., 2022; Acciaio et al., 2024), in particular, in natural language processing, RNNs remain more suitable in a significant range of real-time and online learning tasks that require handling one element of the sequence at a time. The key difference is that transformers are designed to process

entire time sequences at once, using self-attention mechanisms to focus on particular entries of the input, while RNNs use hidden state spaces to retain a memory of previous elements in the input sequence, which makes memory one of the most important features of RNNs. Multiple attempts have been made in recent years to design quantitative measures and characterize memory in neural networks in general (Vershynin, 2020; Koyuncu, 2023) and their recurrent versions, in particular, (Haviv et al., 2019; Li et al., 2021).

The notion of ***memory capacity*** (MC) in recurrent neural networks was first introduced in Jaeger (2002), with a particular focus on the so-called echo state networks (ESNs) (Matthews, 1992; Matthews and Moschytz, 1994; Jaeger and Haas, 2004), which are a popular family of RNNs within the reservoir computing (RC) strand of the literature that have shown to be universal approximants in various contexts (Grigoryeva and Ortega, 2018; Gonon and Ortega, 2020, 2021). RC models are state-space systems whose state map parameters are randomly generated and which can be seen as RNNs with random inner neuron connection weights and a readout layer that is trained depending on the learning task of interest. Memory capacity has been proposed as a measure of the amount of information stored in the states of a state-space system in relation to past inputs. It has been commonly accepted as a valuable metric to evaluate the network's ability to store and extract important information from processed input signals over time. Extensive work has been done in the reservoir computing literature both in the setting of linear (Hermans and Schrauwen, 2010; Dambre et al., 2012; Barancok and Farkas, 2014; Couillet et al., 2016b; Goudarzi et al., 2016; Xue et al., 2017), echo state shallow (White et al., 2004; Farkas et al., 2016; Verzelli et al., 2019), and deep architectures (Gallicchio et al., 2017; Gallicchio, 2018). Memory capacity definitions exploited extensively in the literature are based on a natural observation that the ability of the network to memorize previous inputs can be quantitatively assessed by the correlation between the outputs of the network and its past inputs. Originally, independent and identically distributed input sequences were used for these measurements and only some recent references discuss the case of temporary dependent inputs (for example, Dambre et al. 2012; Charles et al. 2014; Grigoryeva et al. 2016b; Charles et al. 2017; Marzen 2017; Gonon et al. 2020). Proposals of other memory measures have also been discussed in the literature, with Fischer information-based criteria (Ganguli et al., 2008; Tino and Rodan, 2013; Livi et al., 2016; Tino, 2018) among those.

Over the past few years, a series of papers presented analytical expressions for the capacity of time-delay reservoirs (Grigoryeva et al., 2015, 2016a). The main interest of memory measures, in general, and capacities, in particular, is related to their use in architecture design. Once the memory capacity expression as a function of the network (hyper-)parameters is available, one could use it in order to design memory-optimal network architectures. This seemed to be especially important for nontrainable random connectivity neural networks, where the choice of sampling and network structure can be informed by maximizing network capacities. This direction was pursued in numerous studies, with many of those focusing on linear and echo state networks (Ortin et al., 2012; Grigoryeva et al., 2014; Ortín and Pesquera, 2019, 2020).

In this paper, we place ourselves in the setting of linear recurrent neural networks. A recent contribution in the literature in this framework is Gonon et al. (2020). Interestingly, it is proved in this reference that linear systems with white noise inputs (not necessarily independent) and non-singular state autocovariance matrices *automatically have full memory*

capacity, which coincides with the dimension of the state space or, equivalently, the number of neurons in the hidden layer. Moreover, while Jaeger (2002) shows that the memory capacity is maximal if and only if Kalman's controllability rank condition (Kalman, 2010; Sontag, 1991, 1998) is satisfied, Gonon et al. (2020) also proves that the memory capacity of linear systems is given *exactly by the rank of the Kalman controllability matrix*. These results ***contradict numerous studies*** in the literature that report empirical evaluations of the memory capacity of linear recurrent networks inconsistent with the result in Gonon et al. (2020). We shall use the term ***linear memory gap*** to denote the difference between empirically measured memory capacities of linear networks and their provable theoretical values. To the best of our knowledge, this paper is the first to shed light on the nature of this incoherence. We argue that the memory gap originates from pure numerical artifacts overlooked by many previous studies and propose robust techniques that allow for accurate estimation of the memory capacity, which renders full memory results for linear RNNs in agreement with the well-known theoretical results. We claim that multiple efforts in the literature to optimize the memory capacity of linear recurrent networks are hence afflicted by numerical pathologies and convey misleading results.

Specific numerical issues that arise at the time of memory computation and which, as we explain later, are attributed to the ill-conditioning of Krylov matrices, were noticed by some authors in empirical experiments. However, no rigorous explanation has been found so far. Instead, the literature has been developing in the following two directions: the first one designs specific network architectures that are not susceptive to these phenomena, and the second one tunes the hyperparameters to achieve empirical capacity maximization for a given network architecture.

The first research strand finds configurations for which the memory gap is absent or minimal. For example, for ESNs with nonlinear hyperbolic tangent activation, based on empirical insights, Farkas et al. (2016) proposes an orthogonalization process that improves memory capacity evaluation in simulations (similar ideas in the vein of orthogonal neural networks are also developed in White et al. 2004). Strauss et al. (2012) provides designs for ESN reservoir matrices, called RingOfNeurons and ChainOfNeurons, that are inspired by rotation matrices and are based on the memory capacity maximization idea. Full memory of the delay-line and cyclic reservoirs has also been reported in Rodan and Tino (2011, 2012). Tino and Rodan (2013) contribute to the same direction characterizing the MC for a particular type of reservoir connectivity, namely symmetric reservoir matrices. In this paper, we rigorously show how and why particular choices of connectivity matrices in the linear setting surpass the ill-conditioning problem and hence exhibit no memory gap by construction.

The second strand of the literature focuses on the question of whether some hyperparameter choices may maximize the memory capacity of the network. We find that many of these contributions propose explanations of the empirical findings that, in the light of Gonon et al. (2020), are not always entirely correct. In particular, they focus on hyperparameters or sampling distributions of the state map matrix parameters (within the family of regular laws) that have provably no effect on the memory capacity. For example, Gallicchio (2020) makes a numerical argument for the sparsity in ESN reservoir matrices, since it claims that it maximizes memory and the "effective dimension" of the state space. This claim is based on a numerical artifact that is mainly due to the different spectral properties of random

matrix ensembles of different sparsity degrees. Another example is Aceituno et al. (2020) which studies the average eigenvalue modulus of the reservoir matrix as a proxy for memory and suggests, in particular, using circulant matrices to maximize memory.

We conclude this brief literature review by mentioning a few references, which, in our view, are the closest to obtaining a satisfactory explanation of the memory gap phenomenon. Whiteaker and Gerstoft (2022b) correctly identifies the importance of the controllability matrix rank, even though it considers a nonlinear setting. Some intuitive links between the rank of the controllability matrix and memory capacity are discussed in Verzelli et al. (2021) and Whiteaker and Gerstoft (2022a). MC is studied in that paper through simulations that yield an incorrect conclusion as to the imperfect memory in some linear ESN (LESN) designs. Finally, Hermans and Schrauwen (2010) studies memory in the case of continuous-time models and contains a version of the result of input mask neutrality that we present later in the paper.

This paper contains two main contributions. First, we address the methods of empirical memory estimation commonly exploited in the literature. In particular, while Gonon et al. (2020) together with Grigoryeva et al. (2023) prove that N-dimensional linear state-space or RNN systems with randomly sampled matrix parameters of the state map have almost surely full memory of N, both Monte Carlo simulation and algebraic techniques, which we call *naïve*, exhibit numerical issues and lead either to over- or underestimation of the memory capacity in this setting. Unfortunately, these approaches were followed in many studies and led to the devising of recommendations for optimal random architectures based on numerical pathologies, which we discussed in previous paragraphs. Second, the insight into numerical issues led us to develop numerically robust algorithms for memory capacity evaluations. One of the results that we derive is the so-called neutrality of memory capacity to input mask, which we build upon in order to propose a numerically stable memory capacity empirical evaluation scheme using subspace methods. We call these newly introduced techniques the ***orthogonalized subspace*** and ***averaged orthogonalized subspace*** methods. We hope with our proposal to give closure to a long line of contributions in the literature that attempts to maximize capacities that are almost surely and provably full, to begin with.

The paper is structured as follows. In Section 2 we present memory capacity estimation approaches as currently used in the literature of reservoir computing and, specifically, ESN models. We highlight the main issues that arise with these methods, which have lead to significant efforts to seemingly maximize memory properties of linear echo state networks. In Section 3 we present our new method based on linear subspaces induced by the Krylov structure of the controllability matrix. This approach recovers the theoretical memory properties of LESNs and is immediate to implement; we also proposed an improved version that relies on a novel result of the invariance of memory capacity with respect to the choice of the input mask. We conclude with Section 5.

1.1 Code

All codes necessary to reproduce numerical results presented in the paper are publicly available at `https://github.com/Learning-of-Dynamic-Processes/memorycapacity`.

1.2 Notation

Column vectors are denoted by bold lowercase symbols like $\mathbf{r}$. Given a vector $\mathbf{v} \in \mathbb{K}^n$, we denote its entries by v_i, with $i \in \{1, \ldots, n\}$. We denote by $\mathbb{M}_{n,m}$ the space of $\mathbb{K}$-valued $n \times m$ matrices with $m, n \in \mathbb{N}$. The choice of $\mathbb{K}$ is either $\mathbb{C}$ or $\mathbb{R}$, which will be clear from the context. When $n = m$, we use the symbol $\mathbb{M}_n$ to refer to the space of square matrices of order n. Given a vector $\mathbf{v} \in \mathbb{K}^n$, we denote by $\mathrm{diag}(\mathbf{v})$ the diagonal matrix in $\mathbb{M}_n$ with the elements of $\mathbf{v}$ as diagonal entries. Given a matrix $A \in \mathbb{M}_{n,m}$, we denote its components by A_{ij} and we write $A = (A_{ij})$, with $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots m\}$. Given a vector $\mathbf{v} \in \mathbb{R}^n$, the symbol $\|\mathbf{v}\|$ stands for its Euclidean norm. For any $A \in \mathbb{M}_{n,m}$, $\|A\|$ denotes its matrix norm induced by the Euclidean norms in $\mathbb{K}^m$ and $\mathbb{K}^n$, and satisfies that $\|A\| = \sigma_{\max}(A)$, with $\sigma_{\max}(A)$ the largest singular value of A (Horn and Johnson, 2013). Whenever $\mathbb{K} = \mathbb{C}$, for $A \in \mathbb{M}_{n,m}$, we denote by $A^* \in \mathbb{M}_{m,n}$ its conjugate transpose defined by $(A^*)_{ij} = \overline{A_{ji}}$, where the bar denotes the complex conjugate. For $A \in \mathbb{M}_{n,m}$, $A^\top$ denotes its transpose, while $\mathcal{C}(A) \subset \mathbb{K}^n$ and $\mathcal{C}(A^\top) \subset \mathbb{K}^m$ are its column and row spaces, respectively.

2. Linear Memory Capacity

Consider the linear echo state network (LESN) defined by the following two equations:

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + C\mathbf{z}_t + \boldsymbol{\zeta}, \tag{2.1}$$

$$\mathbf{y}_t = W^\top \mathbf{x}_t, \tag{2.2}$$

for $t \in \mathbb{Z}_-$, where $\mathbf{z} \in (\mathbb{R}^d)^{\mathbb{Z}_-}$ are the inputs, $\mathbf{x} \in (\mathbb{R}^N)^{\mathbb{Z}_-}$ are the states, and $\mathbf{y} \in (\mathbb{R}^m)^{\mathbb{Z}_-}$ are the outputs, $d, m, N \in \mathbb{N}$. The states in (2.1) are defined using the ***reservoir (connectivity) matrix*** $A \in \mathbb{M}_N$, the ***input mask*** $C \in \mathbb{M}_{N,d}$, and the ***input shift*** $\boldsymbol{\zeta} \in \mathbb{R}^N$, and are mapped to the outputs via the affine readout map with associated ***readout weights matrix*** $W \in \mathbb{M}_{N,m}$ which can be adjusted to incorporate the intercept term. In the rest of the paper, we consider one-dimensional inputs and outputs and hence use bold symbols $\mathbf{C}, \mathbf{W} \in \mathbb{R}^N$ to denote the input mask and the readouts vectors, respectively.

We shall focus on state-space systems of the type (2.1)-(2.2) that determine an ***input/output*** system. This happens in the presence of the so-called ***echo state property (ESP)***, that is, when for any $\boldsymbol{z} \in \mathbb{R}^{\mathbb{Z}_-}$ there exists a unique $\mathbf{y} \in \mathbb{R}^{\mathbb{Z}_-}$ such that (2.1)-(2.2) hold. One can require that the ESP holds only on the level of the state equation, that is that for any $\mathbf{z} \in (\mathbb{R}^d)^{\mathbb{Z}_-}$ there exists a unique $\mathbf{x} \in (\mathbb{R}^N)^{\mathbb{Z}_-}$ such that (2.1) holds. In Proposition 4.2 in Grigoryeva and Ortega (2021) it is proved that the state equation associated to (2.1) has a unique state-solution $\mathbf{x} \in \ell^\infty_-(\mathbb{R}^N)$ for each input in $\mathbf{z} \in \ell^\infty_-(\mathbb{R})$ (we call this property the $(\ell^\infty_-(\mathbb{R}^N), \ell^\infty_-(\mathbb{R}))$-ESP) if and only if the spectral radius of A is strictly smaller than 1, that is $\rho(A) < 1$. We recall that the inputs $\mathbf{z} \in \ell^\infty_-(\mathbb{R})$ and the inputs $\mathbf{x} \in \ell^\infty_-(\mathbb{R}^N)$ are the left-infinite $\mathbb{R}^N$- and $\mathbb{R}$-valued sequences, respectively, with finite supremum norm $\|\cdot\|_\infty$, that is $\|\mathbf{z}\|_\infty := \sup_{t\in\mathbb{Z}_-}\{|z_t|\} < \infty$ and $\|\mathbf{x}\|_\infty := \sup_{t\in\mathbb{Z}_-}\{\|\mathbf{x}_t\|\} < \infty$ with $\|\cdot\|$ the Euclidean norm. Under the hypothesis $\rho(A) < 1$, the unique solution $\mathbf{x} \in \ell^\infty_-(\mathbb{R}^N)$ of (2.1) associated to the input $\mathbf{z} \in \mathbb{R}^{\mathbb{Z}_-}$ is given by the series

$$\mathbf{x}_t = \sum_{j=0}^{\infty} A^j \mathbf{C} z_{t-j}, \quad t \in \mathbb{Z}_-. \tag{2.3}$$

In this paper, we consider inputs that are realizations of variance-stationary discrete-time stochastic $\mathbb{R}$-valued processes $\mathbf{z} = (z_t)_{t\in\mathbb{Z}_-}$. Additionally, since we study only the memory reconstruction information processing tasks, the target process $\mathbf{y}$ is a forward-shifted version of the input process $\mathbf{z}$. In the stochastic setting, one can show that the same condition $\rho(A) < 1$ is sufficient for the almost sure unique existence of a solution of (2.1). More explicitly, if $\rho(A) < 1$ and the input process $\mathbf{z}$ is such that $\mathrm{Var}(z_t) < c$ for all $t \in \mathbb{Z}_-$ and a finite constant $c > 0$, then there exists an a.s. unique sequence of random variables $\mathbf{x}$ such that

$$\sum_{j=0}^{T} A^j \mathbf{C} z_{t-j} \xrightarrow[T\to\infty]{L^2} \mathbf{x}_t, \quad t \in \mathbb{Z}_-. \tag{2.4}$$

This statement is a corollary of Lütkepohl (2005), Proposition C.9, which requires the absolute summability of the sequence $\left\{A^j\mathbf{C}\right\}_{j\in\mathbb{N}}$ which is, in turn, a consequence of the hypothesis $\rho(A) < 1$ and part (i) of Proposition 4.2 in Grigoryeva and Ortega (2021). Additionally, a proof similar to the one of Proposition 4.1 in Gonon et al. (2020) guarantees that if $\mathbf{z}$ is variance stationary, then so is $\mathbf{x}$. Statements of this type in which the L^2 convergence is replaced by metric convergence in the Wasserstein space can be found in Manjunath and Ortega (2023).

In contrast to conventional recurrent neural networks, where all the network weights (parameters) are subject to training, the parameters of the state equations of reservoir systems are *fixed*, and exclusively the readout map is estimated based on the learning task of interest. More explicitly, within the reservoir computing paradigm, in the case of LESN, the matrix parameters A, $\mathbf{C}$ and $\boldsymbol{\zeta}$ are sampled randomly from (matrix) probability distributions prescribed a priori and $\mathbf{W}$ is estimated. The choice of the law and the properties of these parameters are known to have a significant impact on the performance of the ESN in practical applications.

2.1 Memory Capacity

The notion of memory capacities (MCs) has been introduced in Jaeger (2002) in the context of recurrent neural networks and echo state networks (Matthews, 1992; Matthews and Moschytz, 1994; Jaeger and Haas, 2004) as a way to measure the amount of information contained in the states of a state-space system about the past inputs and to characterize the ability of the network to extract the dynamic features of processed signals. Following Gonon et al. (2020), given a variance-stationary input stochastic process $\mathbf{z} = (z_t)_{t\in\mathbb{Z}_-}$, a state map that satisfies the ESP, and the associated variance-stationary state process $\mathbf{x} = (\mathbf{x}_t)_{t\in\mathbb{Z}_-}$, $\mathbf{x} \in \mathbb{R}^N$, the τ-lag memory capacity of the state-space system with respect to $\mathbf{z}$, with $\tau \in \mathbb{N}$, is defined as

$$\mathrm{MC}_\tau := 1 - \frac{1}{\mathrm{Var}(z_t)} \min_{\mathbf{W}\in\mathbb{R}^N} \mathbb{E}\left[\left(z_{t-\tau} - \mathbf{W}^\top \mathbf{x}_t\right)^2\right], \tag{2.5}$$

where we will often use that $\mathrm{Var}(z_t) = \gamma(0)$ with $\gamma : \mathbb{Z} \mapsto \mathbb{R}$ being the autocovariance function of $\mathbf{z}$.

The ***total memory capacity*** of an ESN is then given by the sum of the capacities at all lags, that is,

$$\mathrm{MC} := \sum_{\tau=0}^{\infty} \mathrm{MC}_\tau. \tag{2.6}$$

It is important to underline that in contrast to what is sometimes defined in the literature (for example, Rodan and Tino 2011), our definition of MC includes lag 0. We are interested in the complete history of the process z_t embedded in the states $\mathbf{x}_t$, including the present. This is consistent with the fact that a LESN where $A = \mathbb{O}_N$ has no memory of inputs at lags $\tau > 0$, but if $N = 1$, it still retains maximal memory as long as $\mathbf{C} \neq \mathbf{0}$. By definition, MC_τ measures how much of the variance of input $z_{t-\tau}$ can be linearly reconstructed from the states $\mathbf{x}_t$. The higher MC_τ is for large τ, the longer the states contain the past history of a sequence of inputs.

Under the assumption that $\Gamma_{\mathbf{x}} := \mathrm{Var}(\mathbf{x}_t)$ is non-singular, MC_τ has the closed-form expression (see Lemma 3.2, Gonon et al. 2020)

$$\mathrm{MC}_\tau = \frac{\mathrm{Cov}(z_{t-\tau}, \mathbf{x}_t)\Gamma_{\mathbf{x}}^{-1}\mathrm{Cov}(\mathbf{x}_t, z_{t-\tau})}{\mathrm{Var}(z_t)}, \ \ \tau \in \mathbb{N}. \tag{2.7}$$

Example 1 (Delay reservoir) Consider the delay (or Takens) reservoir given by the reservoir matrix whose only non-zero elements are $A_{ij} = 1$ for all $j \in \{1, \dots, N-1\}$, $i = j + 1$ (this is usually called a shift matrix), the input mask $\mathbf{C}$ whose all elements are zero except for the first one which is set to one, and the zero input shift $\boldsymbol{\zeta}$. In this case, for any $t \geq N$

$$\boldsymbol{x}_t = \begin{pmatrix} z_t \\ \vdots \\ z_{t-N} \end{pmatrix}$$

and $\mathrm{MC}_\tau = 1$ for $\tau \in \{0, \dots, N\}$ while $\mathrm{MC}_\tau = 0$ for all $\tau \geq N + 1$.

2.1.1 Fischer Memory

An alternative concept developed in the literature that pertains to the memory features of recursive neural networks is that of the ***Fischer memory curve*** (FMC). The idea is introduced in Ganguli et al. (2008) and consists in quantifying the impact of small variations on the current state $\boldsymbol{x}_t$. More precisely, assume that in (2.1) the states are perturbed by i.i.d. noise $(\boldsymbol{\epsilon}_t)_{t\in\mathbb{Z}_-}$ and that $\boldsymbol{\zeta} = \mathbf{0}$. The state equation then reads

$$\boldsymbol{x}_t = A\boldsymbol{x}_{t-1} + \mathbf{C}z_t + \boldsymbol{\epsilon}_t.$$

The Fischer memory matrix is given by

$$\mathrm{F}_{i,j}((z_t)_{t\in\mathbb{Z}_-}) := -\mathbb{E}_{p(\mathbf{x}_t|(z_t)_{t\in\mathbb{Z}_-})}\left[\frac{\partial^2 \log\left(p(\mathbf{x}_t|(z_t)_{t\in\mathbb{Z}_-})\right)}{\partial z_{t-i+1}\,\partial z_{t-j+1}}\right],$$

where $p(\boldsymbol{x}_t|(z_t)_{t\in\mathbb{Z}_-})$ is the input-conditional state distribution, and the Fischer memory curve is given by its diagonal entries, $\mathrm{F}_\tau \equiv \mathrm{F}_{\tau+1,\tau+1}$ for $\tau \geq 0$. Assuming that $\boldsymbol{\epsilon}_t$, for

all $t \in \mathbb{Z}_-$, are mean-zero Gaussian distributed with variance $\sigma_\epsilon^2 \mathbb{I}_N$, one obtains (see the detailed derivations in Ganguli et al. 2008; Tino and Rodan 2013) that $p(\mathbf{x}_t|(z_t)_{t\in\mathbb{Z}_-})$ is Gaussian with the covariance matrix

$$R_{\mathbf{x}} = \sigma_\epsilon^2 \sum_{j=0}^{\infty} A^j (A^\top)^j,$$

and hence the FMC can be written as

$$\mathrm{F}_\tau = \mathbf{C}^\top (A^\top)^\tau R_{\mathbf{x}}^{-1} A^\tau \mathbf{C}.$$

One may easily notice that this formula does *not* depend on input $\mathbf{z}$ and measures memory based only on the architecture properties of the state-space system.

2.1.2 Relation Between Memory Capacities and Fischer Memory

The relation between these two notions of memory is not straightforward. Theorem 1 in Tino and Rodan (2013) shows that

$$\mathrm{MC}_\tau = \sigma_\epsilon^2 \mathrm{F}_\tau + \mathbf{C}^\top (A^\top)^\tau O^{-1} A^\tau \mathbf{C},$$

where $O = \Gamma_{\mathbf{x}}(R_{\mathbf{x}}/\sigma_\epsilon^2 - \Gamma_{\mathbf{x}})^{-1}\Gamma_{\mathbf{x}} + \Gamma_{\mathbf{x}}$, and that it follows $\mathrm{MC}_\tau > \sigma_\epsilon^2 \mathrm{F}_\tau$ for all $\tau > 0$. Due to the complex properties of matrix O, Tino and Rodan (2013) does not establish further general results while providing explicit calculations of both MC and FMC when A is symmetric or orthonormal. Tino (2018) contains further derivations regarding the asymptotic Fischer memory capacity of particular classes of ESN models.

The reasons why in this paper we focus on memory capacity (2.5) instead of the Fischer memory curve are two-fold. First, FMC measures memory only in the state space, and the observation equation (2.2) does not have any impact on the FMC computation. Our primary interest is to evaluate memory in terms of real-world applications, which inevitably requires studying the effect of the linear projection of states onto targets encoded by $\mathbf{W}$. Second, important theoretical contributions towards analyzing the impact of noise on the statistical properties of the states and the linear reservoir systems, in general, have already been made in Couillet et al. (2016a,b). Finally, we emphasize that in Section 3, we are able to show that MC is neutral to the choice of input mask, which is not the case for Fischer memory. As we explain in the following sections, this fact allows us to develop a particular numerical method that is insensitive to numerical artifacts and yields results fully coherent with theory.

2.2 Linear Models Generically Have Maximal Memory

Memory capacities of echo state networks with independent inputs have been originally analyzed in Jaeger (2002). Already in this work, it was shown that

$$1 \leq \mathrm{MC} \leq N.$$

This statement was extended to more general recurrent neural networks in Gonon et al. (2020), where N is in that case the state space dimension. Two results that have recently

appeared in the literature show that linear echo state networks generically achieve *maximal* memory capacity, that is, for almost all LESNs it holds that MC $= N$. Due to their importance in the sequel of the paper and for the sake of completeness, we collect some of those statements in the following result. The first one is contained in Gonon et al. (2020), Corollary 4.2, and we reproduce it in the next proposition with an illustrative proof that will be useful for some derivations later on.

Proposition 1 (LESN Memory Capacity) *Consider a linear ESN model in* (2.1)-(2.2) *and let* $\boldsymbol{\zeta} = \mathbf{0}$. *Let* A *be diagonalizable and such that* $\rho(A) < 1$, *with* $\rho(A)$ *the spectral radius of the matrix* A. *Suppose that all the eigenvalues of* A *are distinct. Let any of the following equivalent conditions hold*

(i) *The vectors* $\{A\mathbf{C}, A^2\mathbf{C}, \ldots, A^N\mathbf{C}\}$ *form a basis of* $\mathbb{R}^N$.

(ii) *The Kalman controllability condition holds.*

(iii) A *has full rank and* $\mathbf{C}$ *is neither the zero vector nor an eigenvector of* A.

If $(z_t)_{t\in\mathbb{Z}_-}$ *is a weakly stationary white noise process, then MC* $= N$.

Proof Under the assumption $\rho(A) < 1$ and of the stationarity and the finite second-order moments of $\mathbf{z}$, the statement (2.4) guarantees that the expression

$$\mathbf{x}_t = \sum_{j=0}^{\infty} A^j \mathbf{C}\, z_{t-j},$$

determines almost surely a unique second-order stationary process. If $\mathrm{Var}(z_t) = \gamma(0)$, the (time-independent) second moment of the state process is

$$\Gamma_{\mathbf{x}} = \gamma(0) \sum_{j=0}^{\infty} A^j \mathbf{C}\, \mathbf{C}^\top (A^j)^\top \tag{2.8}$$

and the covariance of the state and the input process is

$$\mathrm{Cov}(\mathbf{x}_t, z_{t-\tau}) = A^\tau \mathbf{C}\, \gamma(0).$$

Substituting these expressions into (2.7), we conclude that the τ-lag memory capacity is

$$\mathrm{MC}_\tau = \mathbf{C}^\top (A^\top)^\tau \left[\sum_{j=0}^{\infty} A^j \mathbf{C}\mathbf{C}^\top (A^j)^\top\right]^{-1} A^\tau \mathbf{C}, \;\; \tau \in \mathbb{N}, \tag{2.9}$$

where the inverse is well-defined whenever any of the conditions **(i)**, **(ii)**, or **(iii)** is satisfied (see Proposition 4.3, Gonon et al. (2020)). Summing over all lags and using the fact that MC_τ is a scalar, for all $\tau \in \mathbb{N}$, yields

$$\mathrm{MC} = \sum_{\tau=0}^{\infty} \mathbf{C}^\top (A^\top)^\tau \left[\sum_{j=0}^{\infty} A^j \mathbf{C}\mathbf{C}^\top (A^j)^\top\right]^{-1} A^\tau \mathbf{C}$$

$$= \sum_{\tau=0}^{\infty} \operatorname{tr}\left(\left[\sum_{j=0}^{\infty} A^j \mathbf{C}\mathbf{C}^\top (A^\top)^j\right]^{-1} A^\tau \mathbf{C}\mathbf{C}^\top (A^\tau)^\top\right)$$
$$= \operatorname{tr}\left(\left[\sum_{j=0}^{\infty} A^j \mathbf{C}\mathbf{C}^\top (A^j)^\top\right]^{-1} \sum_{\tau=0}^{\infty} A^\tau \mathbf{C}\mathbf{C}^\top (A^\tau)^\top\right) = \operatorname{tr}(\mathbb{I}_N) = N,$$

as required. ■

The second important result was originally presented in Grigoryeva et al. (2023)[1] and guarantees that the conditions **(i)-(iii)** in Proposition 1 are satisfied almost surely, whenever, as it is customary in reservoir computing, the connectivity matrix A and the input mask $\mathbf{C}$ of the linear system (2.1)-(2.2) are randomly drawn from some regular probability distribution. We recall that a random variable $X : \Omega \to \mathbb{R}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and with values on a Borel measurable space $\mathbb{R}$ is *regular* whenever $\mathbb{P}(X = a) = 0$ for all $a \in \mathbb{R}$. The result is stated in the following proposition.

Proposition 2 (Grigoryeva et al. (2023)) *Let $N \in \mathbb{N}$, $A \in \mathbb{M}_N$, and $\mathbf{C} \in \mathbb{R}^N$ and assume that the entries of A and $\mathbf{C}$ are drawn using independent regular real-valued distributions. Then the following statements hold:*

(i) *The vectors $\{\mathbf{C}, A\mathbf{C}, A^2\mathbf{C}, \ldots, A^{N-1}\mathbf{C}\}$ are linearly independent almost surely.*

(ii) *Given m distinct complex numbers $\lambda_1, \ldots, \lambda_m \in \mathbb{C}$, where $m \leq N$, the event that $1, \lambda_1, ..., \lambda_m \notin \sigma(A)$ ($\sigma(A)$ is the spectrum of A) and that the vectors*

$$(\mathbb{I} - \lambda_j A)^{-1}(\mathbb{I} - A)^{-1}(\mathbb{I} - A^N)\mathbf{C}, \quad j = 1, \ldots, m$$

are linearly independent holds almost surely.

Proposition 1 (see also Corollary 4.2 in Gonon et al. 2020) together with Proposition 2 give a definite answer to the question of whether some reservoir architectures have theoretically more memory capacity than others in the linear setting. In theory, as we just showed, all linear ESNs with the connectivity matrix A and the input mask $\mathbf{C}$ drawn from regular distributions *achieve almost surely their upper memory bound* regardless of the underlying design under minimal algebraic conditions. This means that the discussions about optimizing the components of a LESN reservoir's state map to achieve maximal theoretical memory capacity are not justified. Regardless of the LESN architecture, in the setting of Proposition 1, *the memory of any LESN is generically maximal.*

However, empirical estimates in the literature of the memory capacity in applied memory tasks may differ from the theoretical value of N, which has motivated multiple studies with attempts to design LESNs that render "maximized" memory. In the following sections, we characterize the problems associated with the most common ways of numerical estimation of MCs and explain computational issues that yield misleading empirical results. We show that

1. The authors of Grigoryeva et al. (2023) acknowledge that the proof of the proposition has been communicated to them by Friedrich Philipp.

purely numerical pathologies in empirical MC evaluation emerge in a plethora of memory "maximization" techniques applied to LESN architectures. As a solution, we shall propose a simple numerical scheme to combat the numerical inconsistency of empirical estimates with the theoretical result in Proposition 1.

2.3 Monte Carlo Estimation of Memory Capacities

In this section, we address important issues that arise when estimating memory capacities using standard Monte Carlo simulation tools. The definitions and the results discussed in this section show that even in the simplified setting of the so-called regular linear systems, the simulation-based estimation of network capacities may be misleading. We use this section exclusively to motivate the necessity of designing other numerical methods for capacity estimation that do not suffer from the poor statistical properties of naïve approaches based on plug-in estimators. In the following paragraphs, we spell out the finite-sample properties of the natural sample estimator of (total) memory capacity and illustrate the limitations of the sample-based approach that may lead to incorrect memory estimates that are incompatible with the generic N-memory capacity of LESNs.

The availability of the closed-form solution (2.7) facilitates the computation of capacities. However, even for linear specifications, the ill-conditioning of the associated covariance matrices of states leads to technical difficulties. Some of those problems can be handled by using equivalent state-space representations. Proposition 2.5 in Gonon et al. (2020) proves that new representations obtained out of linear injective system morphisms leave capacities invariant and hence can be used to produce systems with more technically tractable properties.

Proposition 3 (Standardization of state-space realizations, Gonon et al. 2020) *Consider a state-space system as in* (2.1)-(2.2) *and suppose that* $\rho(\widetilde{A}) < 1$. *Let* $\mathbf{z} : \Omega \longrightarrow \mathbb{R}^{\mathbb{Z}_-}$ *be a stationary mean-zero input process and let* $\widetilde{\mathbf{x}} : \Omega \longrightarrow (\mathbb{R}_N)^{\mathbb{Z}_-}$ *be the associated stationary state process given by* (2.4). *Suppose that the covariance matrix* $\Gamma_{\widetilde{\mathbf{x}}} := \mathrm{Cov}(\widetilde{\mathbf{x}}_t, \widetilde{\mathbf{x}}_t)$ *is non-singular. Then, the map* $f : \mathbb{R}^N \longrightarrow \mathbb{R}^N$ *given by* $f(\widetilde{\mathbf{x}}) := \Gamma_{\widetilde{\mathbf{x}}}^{-1/2}\widetilde{\mathbf{x}}$ *is a system isomorphism between the system* (2.1)-(2.2) *and the one with state map*

$$\widetilde{F}(\mathbf{x}, z) := A\mathbf{x} + \mathbf{C}z \tag{2.10}$$

and readout

$$\widetilde{h}(\mathbf{x}) := \mathbf{W}^\top \mathbf{x}, \tag{2.11}$$

with $A := \Gamma_{\mathbf{x}}^{-1/2}\widetilde{A}\Gamma_{\mathbf{x}}^{1/2}$, $\mathbf{C} := \Gamma_{\mathbf{x}}^{-1/2}\widetilde{\mathbf{C}}$, *and* $\mathbf{W} = \Gamma_{\mathbf{x}}^{-1/2}\widetilde{\mathbf{W}}$. *Moreover, the state process* $\mathbf{x}$ *associated to the system* $\widetilde{F}$ *and the input* $\mathbf{z}$ *is covariance stationary and*

$$\mathrm{E}[\mathbf{x}_t] = \mathbf{0}, \quad \textit{and} \quad \mathrm{Cov}(\mathbf{x}_t, \mathbf{x}_t) = \mathbb{I}_N. \tag{2.12}$$

This result of invariance of memory capacities with respect to the system isomorphism $f(\mathbf{x}) := \Gamma_{\mathbf{x}}^{-1/2}\mathbf{x}$ allows us to work directly with the standardized state-space systems and assume that $\Gamma_{\mathbf{x}} = \mathbb{I}_N$ without loss of generality.

Definition 4 *Let $\mathbf{z}:\Omega\longrightarrow\mathbb{R}^{\mathbb{Z}_-}$, $D\subset\mathbb{R}$, be a variance-stationary input process and let the state map $\widetilde{F}:\mathbb{R}^N\times\mathbb{R}\to\mathbb{R}^N$ be given by $\widetilde{F}(\mathbf{x},z):=A\mathbf{x}+\mathbf{C}z$ with $\rho(A)<1$. We call a system with the state map $\widetilde{F}$ a **regular linear system** whenever the covariance matrix $\Gamma_{\mathbf{x}}$ of the associated covariance-stationary state process $\mathbf{x}:\Omega\longrightarrow(\mathbb{R}^N)^{\mathbb{Z}_-}$ satisfies $\Gamma_{\mathbf{x}}=\mathbb{I}_N$.*

A straightforward approach to estimate the memory capacity of an echo state network is to simulate the mean zero and variance one process $(z_t)_{t=1}^T$, to compute the associated states $(\mathbf{x}_t)_{t=1}^T$ and to use in (2.7) the plug-in sample estimator

$$\widehat{\boldsymbol{\gamma}_{\mathbf{x}z}}(\tau):=\widehat{\mathrm{Cov}}(\mathbf{x}_t,z_{t-\tau})=\frac{1}{T-\tau}\sum_{t=\tau+1}^{T}\mathbf{x}_t\,z_{t-\tau}. \tag{2.13}$$

This leads to the sample memory capacity estimator

$$\widehat{\mathrm{MC}}_\tau:=\widehat{\boldsymbol{\gamma}_{\mathbf{x}z}}(\tau)^\top\widehat{\boldsymbol{\gamma}_{\mathbf{x}z}}(\tau)=\left\|\widehat{\boldsymbol{\gamma}_{\mathbf{x}z}}(\tau)\right\|_2^2, \tag{2.14}$$

that we refer to as the Monte Carlo estimator. Letting N be fixed, under suitable assumptions of stationarity and sufficiently many finite moments, it is well-known that the above sample estimators are consistent and asymptotically normal (see Brockwell and Davis 2006; Hamilton 1994 and Lütkepohl 2005 for the details). These assumptions hold trivially when z_t is sampled as i.i.d. standard Gaussian noise and we show further that $\widehat{\mathrm{MC}}_\tau\xrightarrow{p}\mathrm{MC}_\tau$ as $T\to\infty$ for any fixed τ.

However, if N is growing with T, $\widehat{\mathrm{Cov}}(\mathbf{x}_t,z_{t-\tau})$ may be inconsistent. In practical implementations of echo state network architectures N can be large, of the order of 10^4 or more. Hence, T must also be appropriately chosen for the Monte Carlo approximations to be valid. These considerations imply the necessity to study memory estimators in the high-dimensional time series setting (see e.g. Chen et al. 2013 or Zhang and Wu 2017 for examples of such discussions). We show in the following paragraphs that inaccuracies when numerically evaluating $\widehat{\mathrm{MC}}_\tau$ even when the ratio T/N is small (in practice $T/N<10$ can already be problematic) mean that the estimator (2.14) is a poor approximation of the LESN memory capacity. This issue can be even more significant when one wishes to quantify MC_τ for τ large. We now provide a quick analysis of these phenomena using standard statistical arguments.

Proposition 5 *Let $N\in\mathbb{N}$, $A\in\mathbb{M}_N$, $\mathbf{C}\in\mathbb{R}^N$, and $\boldsymbol{\zeta}=\mathbf{0}$, and suppose that the resulting linear system is regular. Let $(z_t)_{t=1}^T$, $T\in\mathbb{N}$, be mean-zero i.i.d. Gaussian with $Var(z_t)=\gamma(0)=1$ and let $(\mathbf{x}_t)_{t=1}^T$ be the associated states (obtained using a trivial initialization). Then the memory capacity sample estimator for any $\tau\in\mathbb{N}$, $\tau<T$, is given by*

$$\widehat{\mathrm{MC}}_\tau(T)=\left\|\widehat{\boldsymbol{\gamma}_{\mathbf{x}z}}(\tau)\right\|_2^2 \tag{2.15}$$

with $\widehat{\boldsymbol{\gamma}_{\mathbf{x}z}}(\tau)$ as in (2.13) and the total memory capacity estimator on $\tau_{\max}\in\mathbb{N}$ sample of memory capacities is

$$\widehat{\mathrm{MC}}(T)=\frac{1}{\tau_{\max}}\sum_{\tau=0}^{\tau_{\max}-1}\widehat{\mathrm{MC}}_\tau(T). \tag{2.16}$$

These estimators have the following properties:

(i) $\widehat{\mathrm{MC}}_\tau(T)$ *is a biased estimator of* MC_τ *with bias* B_{MC} *given by*

$$B_{MC} := \mathbb{E}[\widehat{\mathrm{MC}}_\tau(T)] - \mathrm{MC}_\tau = \frac{N}{T-\tau} + \frac{2}{T-\tau}\sum_{j=0}^{\tau} \boldsymbol{\gamma}_{\mathbf{x}z}(j)^\top \boldsymbol{\gamma}_{\mathbf{x}z}(2\tau - j), \tag{2.17}$$

which is positive for large τ*.*

(ii) $\widehat{\mathrm{MC}}_\tau(T)$ *is an asymptotically unbiased estimator, that is* $(\mathbb{E}[\widehat{\mathrm{MC}}_\tau(T)] - \mathrm{MC}_\tau) \to 0$ *as* $T \to \infty$*, and a weakly consistent estimator of* MC_τ*, that is* $\widehat{\mathrm{MC}}_\tau(T) \xrightarrow{p} \mathrm{MC}_\tau$ *as* $T \to \infty$*.*

(iii) $\widehat{\mathrm{MC}}(T)$ *is a biased and asymptotically unbiased estimator of* MC*. Moreover, it is weakly consistent, that is* $\widehat{\mathrm{MC}}(T) \xrightarrow{p} \mathrm{MC}$ *with* $\tau_{\max} = O(T)$ *and* $T \to \infty$*.*

Proof

Both (2.15) and (2.16) are immediate consequences of the definition (2.14). To show **(i)**, we use that by Definition 4, $\rho(A) < 1$ and obtain that

$$\begin{aligned}
\mathbb{E}\big[\big\|\widehat{\boldsymbol{\gamma}_{\mathbf{x}z}(\tau)}\big\|_2^2\big] =& \mathbb{E}\left[\left(\frac{1}{T-\tau}\sum_{t=\tau+1}^{T} \mathbf{x}_t\, z_{t-\tau}\right)^\top \left(\frac{1}{T-\tau}\sum_{s=\tau+1}^{T} \mathbf{x}_s\, z_{s-\tau}\right)\right] \\
=& \frac{1}{(T-\tau)^2}\mathbb{E}\left[\left(\sum_{t=\tau+1}^{T}\sum_{j=0}^{\infty} A^j\mathbf{C}\, z_{t-j}\, z_{t-\tau}\right)^\top \left(\sum_{s=\tau+1}^{T}\sum_{k=0}^{\infty} A^k\mathbf{C}\, z_{t-k}\, z_{s-\tau}\right)\right] \\
=& \frac{1}{(T-\tau)^2}\,\mathbb{E}\left[\mathbf{C}^\top\left\{\sum_{t=\tau+1}^{T}\sum_{s=\tau+1}^{T}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}(A^j)^\top A^k z_{t-\tau}\, z_{t-j} z_{s-\tau} z_{s-k}\right\}\mathbf{C}\right] \\
=& \frac{1}{(T-\tau)^2}\Bigg\{\sum_{t=\tau+1}^{T}\mathbf{C}^\top(A^\tau)^\top A^\tau\mathbf{C}\,\mathbb{E}\left[z_{t-\tau}^4\right] + \sum_{t\neq s}\mathbf{C}^\top(A^\tau)^\top A^\tau\mathbf{C}\,\mathbb{E}\left[z_{t-\tau}^2\, z_{s-\tau}^2\right] \\
&+ \sum_{t=\tau+1}^{T}\sum_{j=0,j\neq\tau}^{\infty}\mathbf{C}^\top(A^j)^\top A^j\mathbf{C}\,\mathbb{E}\left[z_{t-\tau}^2\, z_{t-j}^2\right] \\
&+ \sum_{t=\tau+1}^{T}\sum_{j=0,j\neq\tau}^{2\tau}\mathbf{C}^\top(A^j)^\top A^{2\tau-j}\mathbf{C}\,\mathbb{E}\left[z_{t-\tau}^2\, z_{t-j}^2\right]\Bigg\} \\
=& \frac{1}{T-\tau}\big\{3\|\boldsymbol{\gamma}_{\mathbf{x}z}(\tau)\|_2^2 + (T-\tau-1)\|\boldsymbol{\gamma}_{\mathbf{x}z}(\tau)\|_2^2 + \gamma(0)\mathrm{tr}(\Gamma_{\mathbf{x}}) - \|\boldsymbol{\gamma}_{\mathbf{x}z}(\tau)\|_2^2 \\
&+ \sum_{j=0}^{2\tau}\mathbf{C}^\top(A^j)^\top A^{2\tau-j}\mathbf{C}\gamma(0)^2 - \|\boldsymbol{\gamma}_{\mathbf{x}z}(\tau)\|_2^2\big\} \\
=& \|\boldsymbol{\gamma}_{\mathbf{x}z}(\tau)\|_2^2 + \frac{1}{(T-\tau)}\gamma(0)\mathrm{tr}(\Gamma_{\mathbf{x}}) + \frac{1}{(T-\tau)}\sum_{j=0}^{2\tau}\boldsymbol{\gamma}_{\mathbf{x}z}(j)^\top\boldsymbol{\gamma}_{\mathbf{x}z}(2\tau-j),
\end{aligned}$$

where we can use that $\Gamma_{\mathbf{x}} = \mathbb{I}_N$ and that $\gamma(0) = 1$, which yields (2.17).

Further, using that for any $\epsilon > 0$ there exists a matrix norm $|||\cdot|||$ such that $|||A||| = \rho(A)+\epsilon$ (see Lemma 7.6.12 in Horn and Johnson 2013), the second term in B_{MC} can be bounded as follows

$$\begin{aligned}\frac{1}{T-\tau}\sum_{j=0}^{2\tau}|\boldsymbol{\gamma}_{\mathbf{x}z}(j)^\top\boldsymbol{\gamma}_{\mathbf{x}z}(2\tau-j)| &= \frac{\gamma(0)^2}{T-\tau}\sum_{j=0}^{2\tau}|\mathbf{C}^\top(A^j)^\top A^{2\tau-j}\mathbf{C}| \\ &\leq \frac{\gamma(0)^2}{T-\tau}\sum_{j=0}^{2\tau}|||\mathbf{C}|||^2\left|\left|\left|(A^\top)^j\right|\right|\right||A^{2\tau-j}||| \\ &\leq \frac{\gamma(0)^2}{T-\tau}|||\mathbf{C}|||^2\sum_{j=0}^{2\tau}\left|\left|\left|A^\top\right|\right|\right|^j|||A|||^{2\tau-j} \\ &= \frac{\tau+1}{T-\tau}\gamma(0)^2|||\mathbf{C}|||^2(\rho(A)+\epsilon)^{2\tau},\end{aligned}$$

and hence decays exponentially fast with τ. It is also easy to see that B_{MC} is always positive for large enough τ.

In order to show **(ii)**, notice that **(i)** together with the Markov inequality gives $\widehat{\mathrm{MC}}_\tau(T) = O_p(T^{-1})$ which yileds asymptotic unbiasedness as $T\to\infty$ and weak consistency of $\widehat{\mathrm{MC}}_\tau(T)$ as an estimator of MC_τ. Finally, in **(iii)** one can mimic the proof of **(ii)** and use that, by Gelfand's formula (Lax, 2002), $\lim_{k\to\infty}\left|\left|\left|A^k\right|\right|\right|^{1/k} = \rho(A) < 1$, which implies the existence of a number $k_0\in\mathbb{N}$ such that $\left|\left|\left|A^k\right|\right|\right| < 1$, for all $k\geq k_0$. Consequently, this implies the finiteness of all the sums in $\widehat{\mathrm{MC}}_\tau(T)$ with $\tau_{\max} = O(T)$ and $T\to\infty$. ■

This result shows that even though the estimator of the memory capacity is asymptotically unbiased, in finite samples MC_τ is always positively biased above zero. This means that, even with large T summing up $\tau_{\max}$ terms in the sequence $\{\mathrm{MC}_\tau\}_{\tau=0}^{\infty}$ may yield a memory capacity estimate that is above the theoretical limit given by Proposition 1. This happens even when reservoir matrices are well-conditioned. Figure 1 illustrates the case when $N = 100$, A is a scaled random orthogonal matrix, and $\mathbf{C}$ is a 2-norm-scaled random normal vector. Taking $\tau_{\max} = 500$ to estimate MC, we show that even in those Monte Carlo simulations where $T/N \approx 100$ non-negligible memory overestimation errors are committed.

2.4 Naïve Algebraic Memory Estimation

In this section, we consider another possibility for the evaluation of the memory using a purely algebraic approach and without relying on Monte Carlo simulations. Again, we show that numerical issues are also encountered with this approach. We start by noticing that, under the hypotheses in Proposition 1, the memory capacity can be computed using (2.9), namely, for any $\tau\in\mathbb{N}$

$$\mathrm{MC}_\tau = \mathbf{C}^\top(A^\tau)^\top G_{\mathbf{x}}^{-1}A^\tau\mathbf{C}, \tag{2.18}$$

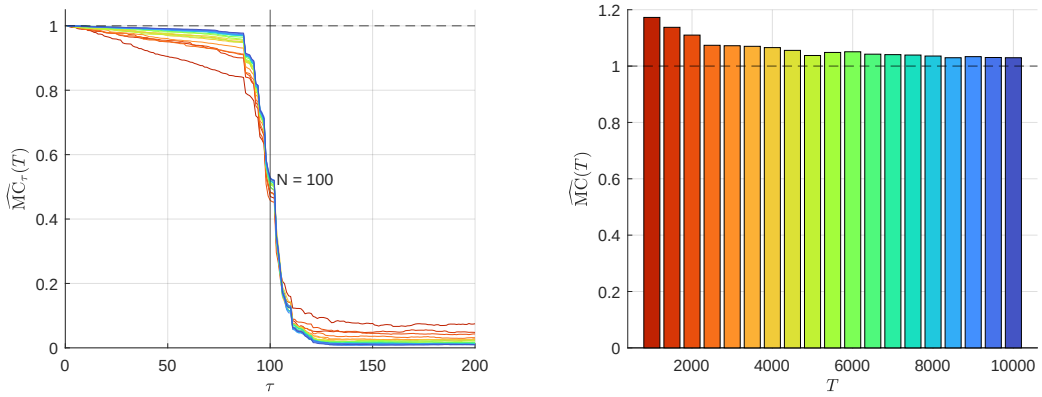


Figure 1: Illustration of memory capacity inflation due to the inconsistent estimation of MC_τ for LESN with $N = 100$, orthogonal A with $\rho(A) = 0.9$, and input mask $\mathbf{C} = \overline{\mathbf{C}}/\|\overline{\mathbf{C}}\|$ with $\mathbf{C} = (\bar{c}_i)_{i=1}^N \sim$ i.i.d. $\mathcal{N}(0,1)$: (a) memory curves $\widehat{\mathrm{MC}}_\tau(T)$; (b) bar chart of normalized total memory capacity $\widehat{\mathrm{MC}}(T)/N$. Memory curves $\widehat{\mathrm{MC}}_\tau(T)$ are computed for $\tau \in \{0, 1, ..., 5N\}$ (in (a), $\widehat{\mathrm{MC}}_\tau(T)$ is plotted only up to $\tau = 2N$ for the sake of clarity). Estimators are computed from simulated $(z_t)_{t=1}^T \sim$ i.i.d. $\mathcal{N}(0,1)$, with $T \in \{1000, 1500, \ldots, 10000\}$.

where $G_{\mathbf{x}} := \gamma(0)^{-1}\Gamma_{\mathbf{x}}$ denotes the normalized version of the state autocovariance matrix in (2.8) and can be written as

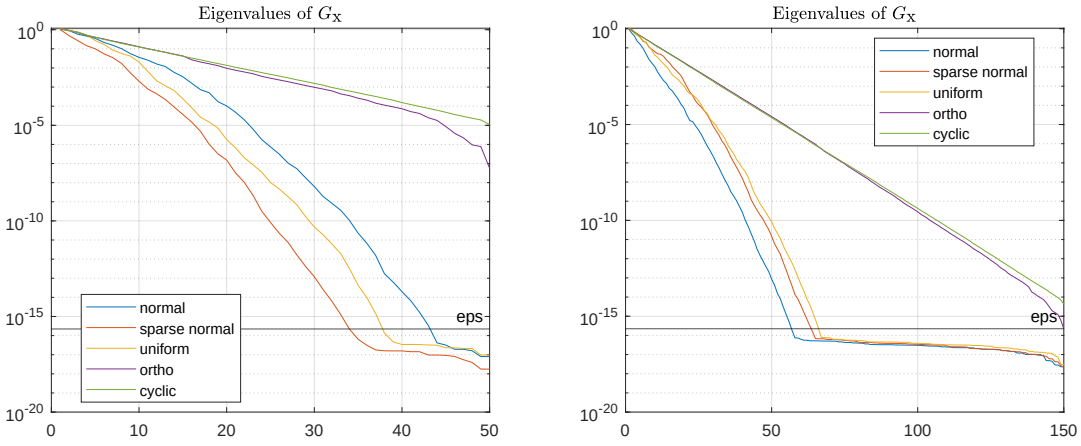
$$G_{\mathbf{x}} = \sum_{j=0}^{\infty} A^j \mathbf{C}\mathbf{C}^\top (A^j)^\top. \tag{2.19}$$

The infinite series in the definition of $G_{\mathbf{x}}$ may be hard to approximate well with a finite number of terms if the spectral radius of A is very close to one, a choice that is quite common in applications. This concern can be easily mitigated by noting that under the hypotheses of Proposition 1 a closed-form expression of $G_{\mathbf{x}}$ in terms of the eigendecomposition of A can be derived (for details, we refer the reader to the proof of Proposition 4.3 in Gonon et al. (2020)). Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$ be an eigenbasis of A and $\{\lambda_1, \ldots, \lambda_N\}$ be the associated eigenvalues. By expressing $\mathbf{C}$ as $\mathbf{C} = \sum_{i=1}^N c_i \mathbf{v}_i$ it is straightforward to show that

$$G_{\mathbf{x}} = \sum_{i,j=1}^{N} \frac{c_i \bar{c}_j}{1 - \lambda_i \bar{\lambda}_j} \mathbf{v}_i \mathbf{v}_j^*. \tag{2.20}$$

and hence $G_{\mathbf{x}}$ can be readily and precisely computed. Unfortunately, $G_{\mathbf{x}}$ can still be significantly poorly conditioned for moderately large N and commonly chosen distributions for the entries of A. This problem is easy to illustrate by plotting the norm of the eigenvalues of $G_{\mathbf{x}}$ for A sampled from laws that are standard in the literature.

We provide an example demonstrating this phenomenon in Figure 2 using two reservoirs of size $N = 50$ and $N = 150$. More precisely, for five commonly used choices of connectivity matrices, we plot the absolute values of the eigenvalues of $G_{\mathbf{x}}$ in decreasing order. We compare them with the standard double-precision of floating point numbers *eps* in our



(a) $N = 50$ (b) $N = 150$

Figure 2: Eigenvalue plot (in absolute values) for $G_{\mathbf{x}}$ for various types of connectivity matrices. $G_{\mathbf{x}}$ was computed using 1000 series terms in (2.18), a connectivity matrix $A \in \mathbb{M}_N$ with spectral radius $\rho(A) = 0.9$ and a unit norm input mask $\mathbf{C} \in \mathbb{R}^N$. Computations are performed in MATLAB with the standard double-precision of floating point numbers $eps = 2^{-52} \approx 2.2 \times 10^{-16}$ marked with the black horizontal solid line.

software of choice, MATLAB. Notice that all the eigenvalues in absolute value smaller than *eps* will be numerically treated as zero by linear algebra routines. This poor conditioning does not by itself mean that software packages will fail to solve the linear system given by $G_{\mathbf{x}}\,\mathbf{u} = A^{\tau}\mathbf{C}$; rather, the numerical solution for $\mathbf{u}$ will be inaccurate (Horn and Johnson, 2013, Section 5.8). This numerical instability is at the origin of the seemingly suboptimal memory performance of LESNs observed in implementations. Further, note that even if $\Gamma_{\mathbf{x}}$ is estimated using a Monte Carlo simulation, due to its consistency as $T \to \infty$, the sample estimator $\widehat{\Gamma}_{\mathbf{x}}$ inherits the conditioning issues of its theoretical counterpart. This effectively implies that the simulation of $G_{\mathbf{x}}$ is not a feasible way to mitigate the conditioning problem, even asymptotically. Regularization methods, such as Tikhonov, also do not solve this as they modify the covariance eigenvalue structure.

The following example about the so-called *cyclic reservoirs* is much studied in the literature under the name "RingOfNeurons" (see Strauss et al. 2012, for example, or more recently in Verzelli et al. 2021). Cyclic architectures yield memory curves that are computable in closed form, and the ill-conditioning of $G_{\mathbf{x}}$ can be explicitly demonstrated. Cyclic ESNs fall into the more general category of orthogonal recurrent neural networks, for which White et al. (2004) has also derived some theoretical memory properties. However, there the authors consider the case in which the states may be contaminated by noise, a situation that we do not discuss in this work.

Example 2 (Cyclic reservoirs) Consider a N-dimensional cyclic reservoir with the unscaled orthogonal connectivity matrix

$$\widetilde{A} = \begin{pmatrix} 0 & 0 & \dots & 0 & 1 \\ 1 & 0 & \ddots & 0 & 0 \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & 0 \\ 0 & \dots & 0 & 1 & 0 \end{pmatrix} \in \mathbb{M}_N,$$

which is rescaled with some $\rho_A < 1$ by setting $A = \rho_A \widetilde{A}$.

In the literature, both $\widetilde{A}$ and $\widetilde{A}^\top$ are referred to as the cyclic reservoir matrices, and the state dynamics they define are identical up to a permutation of the reservoir nodes (Rodan and Tino, 2011). Let $\mathbf{C} = \mathbf{e}_1$ be the first canonical basis vector of $\mathbb{R}^N$. First, observe that

$$A\mathbf{C} = \mathbf{e}_2, \quad A^2\mathbf{C} = \mathbf{e}_3, \, \dots, \, A^{N-1}\mathbf{C} = \mathbf{e}_N,$$

which justifies the use of the term *cyclic*.[2] Second, note that in this case we can obtain the explicit expression of the normalized state covariance matrix as follows:

$$G_{\mathbf{x}} = \operatorname{diag}\left(\sum_{j=0}^{\infty} \rho_A^{i(2N)}, \sum_{j=0}^{\infty} \rho_A^{i(2N)+2}, \dots, \sum_{j=0}^{\infty} \rho_A^{i(2N)+2(N-1)} \right) = \operatorname{diag}\left(\frac{1}{1-\rho_A^{2N}}, \frac{\rho_A^2}{1-\rho_A^{2N}}, \dots, \frac{\rho_A^{2(N-1)}}{1-\rho_A^{2N}} \right)$$

and hence

$$G_{\mathbf{x}}^{-1} = \operatorname{diag}\left(1-\rho_A^{2N}, \; \frac{1-\rho_A^{2N}}{\rho_A^2}, \; \dots \; , \; \frac{1-\rho_A^{2N}}{\rho_A^{2(N-1)}} \right).$$

This formula shows that if N is large, inversion of $G_{\mathbf{x}}$ can be an ill-conditioned problem depending on ρ_A. Finally, for $0 \le \tau \le N-1$ it holds

$$\mathrm{MC}_\tau = \mathbf{e}_1^\top (A^\tau)^\top G_{\mathbf{x}}^{-1} A^\tau \mathbf{e}_1 = \rho_A^\tau \left(\frac{1-\rho_A^{2N}}{\rho_A^{2\tau}} \right) \rho_A^\tau = 1 - \rho_A^{2N},$$

while in general for $kN \le \tau \le k(N+1)-1$, $k > 1$, one has

$$\mathrm{MC}_\tau = \rho_A^{kN+\tau} \left(\frac{1-\rho_A^{2N}}{\rho_A^{2\tau}} \right) \rho_A^{kN+\tau} = \rho_A^{2kN}(1-\rho_A^{2N}).$$

These computations are a special case of more general results in Rodan and Tino (2011), although we have made explicit the values of MC_τ. Rodan and Tino (2011) further proved that such memory capacities arise for generic $\mathbf{C}$ when A is chosen to be a regular rotation based on the input mask.

2. Further, in Proposition 6 we prove that memory capacities MC_τ are invariant with respect to the choice of $\mathbf{C}$. Hence, our selection of input mask does not imply any loss of generality.

3. Robust Memory Computation

In this section, we propose simple but effective methods to compute the memory capacity MC_τ for linear ESNs. These methods are not affected by the problems discussed in Section 2. First, we show a strong neutrality result of the memory capacity with respect to input masks. Second, we discuss the origin of numerical instabilities of memory computation and the so-called memory gaps, borrowing from the theory of Krylov subspaces. Finally, we propose new computational methods based on the Arnoldi iteration algorithm for the leading eigenvector computation and on the memory neutrality with respect to the input mask. We call our proposed methods *robust*, since they do not suffer explicitly from the conditioning issues that arise in the naïve algebraic and statistical methods presented in the previous section and render empirical results that are in agreement with the theory.

3.1 Input Mask Memory Neutrality

A fundamental aspect of memory capacity is its dependence on the structure of the connectivity matrix A and the input mask $\mathbf{C}$. We recall that by Proposition 1, we know that as long as A and $\mathbf{C}$ satisfy a controllability condition, the total memory MC of a LESN is maximal. Moreover, by Proposition 2 this holds almost surely whenever both A and $\mathbf{C}$ are sampled from some regular distribution. We now prove a much stronger result: in the linear setup, under the same controllability conditions, the input mask $\mathbf{C}$ does not have any impact on individual τ-lag memory capacities MC_τ.

Proposition 6 (Input mask neutrality) *For any linear echo state network under the assumptions of Proposition 1, the memory capacity is input mask neutral, that is,* MC_τ *is invariant with respect to the choice of* $\mathbf{C}$*, for all* $\tau \in \mathbb{N}$*.*

Proof Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$ be an eigenbasis of A and $\{\lambda_1, \ldots, \lambda_N\}$ be the associated eigenvalues. Denote $\Lambda := \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$, $V := (\mathbf{v}_1 | \mathbf{v}_2 | \ldots | \mathbf{v}_N)$, and

$$V^{-1} = \begin{pmatrix} \mathbf{v}_1^* \\ \vdots \\ \mathbf{v}_N^* \end{pmatrix},$$

and notice that by the hypothesis of diagonalizability of A one has $A = V\Lambda V^{-1}$. Using the eigenbasis of A, or using the columns of V, it holds for the input mask that $\mathbf{C} = \sum_{i=1}^N c_i \, \mathbf{v}_i$ with $\mathbf{c} := (c_1, \ldots, c_N)^\top$ the vector of coefficients. We now recall that by (2.20)

$$G_{\mathbf{x}} = \sum_{j=0}^{\infty} A^j \mathbf{C}\mathbf{C}^\top (A^j)^\top = \sum_{i,j=1}^{N} \varphi_{i,j} \, \mathbf{v}_i \, \mathbf{v}_j^*,$$

with $\varphi_{i,j} := (c_i \bar{c}_j)/(1 - \lambda_i \overline{\lambda}_j)$, and hence it holds that

$$V^{-1} G_{\mathbf{x}} (V^*)^{-1} = \left(\sum_{i,j=1}^{N} \varphi_{i,j} \left(\mathbf{v}_k^* \, \mathbf{v}_i \, \mathbf{v}_j^* \, \mathbf{v}_l \right) \right)_{k,l}^{N} = (\varphi_{k,l})_{k,l}^{N} .$$

Finally, using this expression in (2.18), we can write MC_τ as follows:

$$\begin{aligned}
\mathrm{MC}_\tau &= \mathbf{C}^\top (A^\tau)^\top G_{\mathbf{x}}^{-1} A^\tau \mathbf{C} = \mathbf{C}^\top (V^*)^{-1} (\Lambda^*)^\tau V^* G_{\mathbf{x}}^{-1} V \Lambda^\tau V^{-1} \mathbf{C} \\
&= \mathbf{C}^\top (V^{-1})^* (\Lambda^*)^\tau \left((\varphi_{k,l})_{k,l}^N \right)^{-1} \Lambda^\tau V^{-1} \mathbf{C} = \mathbf{c}^* (\Lambda^*)^\tau \left((\varphi_{k,l})_{k,l}^N \right)^{-1} \Lambda^\tau \mathbf{c} \\
&= \mathbf{c}^* (\Lambda^*)^\tau \left(\mathrm{diag}\,(\mathbf{c}) \left(\frac{1}{1 - \lambda_k \overline{\lambda}_l} \right)_{k,l}^N \mathrm{diag}\,(\mathbf{c}^*) \right)^{-1} \Lambda^\tau \mathbf{c} \\
&= \mathbf{c}^* (\Lambda^*)^\tau \mathrm{diag}\,(\mathbf{c}^*)^{-1} \left(\left(\frac{1}{1 - \lambda_k \overline{\lambda}_l} \right)_{k,l}^N \right)^{-1} \mathrm{diag}\,(\mathbf{c})^{-1} \Lambda^\tau \mathbf{c} \\
&= \boldsymbol{\iota}_N^\top (\Lambda^*)^\tau \left(\left(\frac{1}{1 - \lambda_k \overline{\lambda}_l} \right)_{k,l}^N \right)^{-1} \Lambda^\tau \boldsymbol{\iota}_N, \qquad (3.1)
\end{aligned}$$

where $\boldsymbol{\iota}_N = (1, \ldots, 1)^\top \in \mathbb{R}^N$. The last equality in the derivation follows from the commutative property of the product of diagonal matrices. Hence, MC_τ is independent of $\mathbf{C}$ for all $\tau \in \mathbb{N}$ under the stated assumptions. ■

A complementary result in continuous time with stationary inputs was derived by Hermans and Schrauwen (2010). To the best of our knowledge, the previous proposition is the first derivation of this property in the context of discrete-time models. A generalization of the memory neutrality for weakly stationary inputs (possibly autocorrelated) is given in Theorem 9 in Appendix A.

3.1.1 Another Formula for Memory Capacity

The proof of Proposition 6 offers another additional strategy that one may follow in order to compute memory capacities. Indeed, the resulting closed-form expression (3.1) can be used to evaluate the memory curve. More precisely, for a chosen reservoir matrix A it is sufficient to compute its eigendecomposition $A = V \Lambda V^{-1}$, then construct the matrix

$$L_A := \left(\frac{1}{1 - \lambda_k \overline{\lambda}_l} \right)_{k,l}^N,$$

and finally compute

$$\mathrm{MC}_\tau = \boldsymbol{\iota}_N^\top (\Lambda^*)^\tau L_A^{-1} \Lambda^\tau \boldsymbol{\iota}_N.$$

Unfortunately, similarly to all the previous approaches, this strategy still exploits the structure of the spectrum of A and may suffer from the same ill-conditioning issues. Simple simulations, which, for the sake of brevity, we do not report, immediately show that regular matrix distributions produce L_A matrices with eigenvalues decaying as quickly as those of the respective $G_{\mathbf{x}}$. This makes the direct application of Proposition 6 for memory evaluation also an infeasible option.

Despite the fact that the result of the neutrality of the LESN memory with respect to the choice of the input mask in Proposition 6 yields no immediate numerical advantages, it is nevertheless at the origin of robust numerical techniques for empirical memory evaluation that we present in the following sections. More explicitly, we shall show how to use the

memory neutrality property to design a memory capacity estimation procedure that recovers full memory in linear ESN models and is robust with respect to the numerical issues discussed in Section 2.

3.2 Krylov Conditioning

In Section 2.4 we showed that the normalized covariance matrices $G_{\mathbf{x}}$ intervene in the computation of capacities MC_τ, $\tau \in \mathbb{N}$. We now explain how one of the sources of numerical problems in memory capacity evaluation is due to the poor conditioning of Krylov matrices that are implicitly used in numerical procedures when evaluating $G_{\mathbf{x}}$.

For $N \in \mathbb{N}$, $A \in \mathbb{M}_N$, and $\mathbf{C} \in \mathbb{R}^N$ define the Krylov matrix

$$K := \left(\mathbf{C} \,|\, A\mathbf{C} \,|\, A^2\mathbf{C} \,|\, \ldots\right),$$

which is infinite in the column dimension. Under the hypothesis $\rho(A) < 1$, Gelfand's formula (Lax, 2002) guarantees that there exists $k_0 \in \mathbb{N}$ such that $\left|\!\left|\!\left|A^{k_0}\right|\!\right|\!\right|_\infty < 1$ and hence for any $\epsilon > 0$ there exists $k \in \mathbb{N}$ such that $\left|\!\left|\!\left|A^{k}\right|\!\right|\!\right|_\infty < \epsilon$. We can use this fact to truncate the matrix K to m columns so that $\|A^m\mathbf{C}\|_\infty < eps$, with eps denoting the double-precision of floating numbers of the researcher's numerical software. Therefore, when using numerical tools, the finite-dimensional m-column Krylov matrix is used. We denote this matrix by

$$K_m := \left(\mathbf{C} \,|\, A\mathbf{C} \,|\, A^2\mathbf{C} \,|\, \ldots \,|\, A^{m-1}\mathbf{C}\right) \tag{3.2}$$

and notice that $G_{\mathbf{x}}$ can be approximated by the product of finite-dimensional matrices,

$$\widetilde{G}_{\mathbf{x}} = K_m K_m^\top. \tag{3.3}$$

A useful factorization of the finite-dimensional Krylov matrices is given by the following result, which we adapt from Lemma 2.4 in Meurant and Duintjer Tebbens (2020) using our notation.

Lemma 7 *Let $A \in \mathbb{M}_N$ be diagonalizable with $A = V\Lambda V^{-1}$, where matrix Λ is diagonal, and let $\mathbf{c} = V^{-1}\mathbf{C}$. Then, the Krylov matrices K_m defined in* (3.2) *can be factorized as*

$$K_m = V\, D_{\mathbf{c}}\, W_m, \tag{3.4}$$

where $D_{\mathbf{c}} = \mathrm{diag}(\mathbf{c})$ and $W_m \in \mathbb{M}_{N,m}$ is the Vandermonde matrix

$$W_m := \begin{pmatrix} 1 & \lambda_1 & \cdots & \lambda_1^{m-1} \\ 1 & \lambda_2 & \cdots & \lambda_1^{m-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_N & \cdots & \lambda_N^{m-1} \end{pmatrix}$$

constructed using the eigenvalues of A.

It is well-known that Krylov matrices are difficult to treat numerically. As pointed out in Meurant and Duintjer Tebbens (2020), K_m is often lacking in numerical rank when compared to the theoretical rank N, and, more importantly, it can have exponentially

increasing conditioning number as m grows. In our case, this phenomenon can be observed by noting that the eigenvalues of $\widetilde{G}_{\mathbf{x}} = K_m K_m^\top$ are the same as nonzero eigenvalues of $K_m^\top K_m$ for which it holds that

$$K_m^\top K_m = W_m^* \, D_{\mathbf{c}}^* V^* V D_{\mathbf{c}} \, W_m.$$

The right-hand side of this expression, under the assumption that A is normal and $D_{\mathbf{c}} = \mathbb{I}_N$, results in the positive-definite Hankel matrix $W_m^* W_m$. Tyrtyshnikov (1994) proved that for real positive-definite Hankel matrices and general Krylov matrices the spectral condition number has exponential lower bounds in m, which means that $\widetilde{G}_{\mathbf{x}}$ can indeed be extremely ill-conditioned in many common setups.

3.3 Memory Gaps and Krylov Subspace Squeezing

As we already pointed out several times, Theorem 4.4 in Gonon et al. (2020) states that the total memory capacity MC of a LESN equals

$$\mathrm{MC} = \mathrm{rank}\{K_N\},$$

with $K_N \in \mathbb{M}_N$ as in (3.2). We refer to the discrepancy between this theoretical result and its numerical estimation as ***memory gap***. The next paragraphs propose an explanation of why so often there is a disagreement between theoretical and empirically computed memory capacities.

3.3.1 Geometric Interpretation of Krylov Subspace Squeezing

We start by introducing the Krylov subspaces and their squeezing, which results in memory gaps in empirical exercises. We refer the reader to some interesting literature regarding the theory of Krylov subspace methods (Bellalij et al., 2016), its geometric aspects (Eiermann and Ernst, 2001), and use for linear (Meurant and Duintjer Tebbens, 2020) and nonlinear systems (Hashimoto et al., 2020).

Definition 8 (Krylov subspace) *The* ***jth-order Krylov subspace*** *generated by a matrix $A \in \mathbb{M}_N$ and a vector $\mathbf{C} \in \mathbb{R}^N$ is the linear subspace of $\mathbb{R}^N$ given by*

$$\mathcal{K}_j(A, \mathbf{C}) = \mathrm{span}\left\{\mathbf{C}, A\mathbf{C}, A^2\mathbf{C}, \ldots, A^{j-1}\mathbf{C}\right\}.$$

Let now N be large and consider the QR decomposition of the Krylov matrix

$$K_N = (\mathbf{q}_1|\mathbf{q}_2|\ldots|\mathbf{q}_N) \begin{pmatrix} r_{1,1} & r_{1,2} & \ldots & r_{1,N} \\ 0 & r_{2,2} & \ldots & r_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & r_{N,N} \end{pmatrix} = QR.$$

If Q and R are obtained via Gram-Schmidt orthogonalization, then the diagonal entries of R have a clear geometric interpretation: each $r_{j,j}$ represents the norm of the orthogonal component in vector $A^j\mathbf{C}$ with respect to the subspace spanned by the columns of the

Krylov matrix K_j, or equivalently $\mathcal{K}_j(A, \mathbf{C})$. Here, we ignore the fact that the Gram-Schmidt implementation of QR is numerically unstable and instead focus on the fact that matrix R inherits the rank structure of K_N.

In practice, we observe that the size of $r_{j,j}$ decays *superexponentially* compared to the decay of powers of $\rho(A)$, a phenomenon that we term ***Krylov subspace squeezing***. This means that for A with $\rho(A) < 1$ and with large enough N there exists a positive integer $\ell < N$ such that numerically

$$R \approx \begin{pmatrix} R_1 & R_2 \\ \mathbb{O}_{N-\ell,\ell} & \mathbb{O}_{N-\ell,N-\ell} \end{pmatrix}.$$

This implies that naïve methods, which do not control for the ill-conditioning of $G_{\mathbf{x}}$, lead to the incorrect estimate

$$\mathrm{MC} = \mathrm{rank}\{R\} \approx \ell.$$

We construct a simulation to showcase the Krylov subspace squeezing phenomenon for commonly chosen distributions of the reservoir connectivity matrix A.

Let $j \in \mathbb{N}$ and denote as $\boldsymbol{\theta}_{j+1} = \mathrm{perp}_{\mathcal{K}_j(A,\mathbf{C})}(A^j\mathbf{C}) \in \mathcal{K}_j(A, \mathbf{C})^{\perp}$ the orthogonal component of $A^j\mathbf{C}$ with respect to $\mathcal{K}_j(A, \mathbf{C})$, with $\|\boldsymbol{\theta}_1\| = \|\mathbf{C}\|$ and hence $\|\boldsymbol{\theta}_1\| = 1$ due to normalization. To compute $\boldsymbol{\theta}_j$ in a robust fashion, we employ two different approaches. Firstly, one can use the ***Arnoldi iteration approach*** (Arnoldi, 1951), which is specially designed to handle the orthogonalization of Krylov iterations. Alternatively, one can define the projection $P_j^c : \mathbb{R}^N \longrightarrow \mathcal{K}_j(A, \mathbf{C})$, or, equivalently, $P_j^c : \mathbb{R}^N \longrightarrow \mathcal{C}(K_j)$, with the corresponding projection matrix $P_j^c = K_j(K_j^\top K_j)^{-1}K_j^\top$. Additionally, we may take the singular value decomposition of K_j given by

$$K_j = U_j\, \Sigma_j\, W_j^\top, \tag{3.5}$$

where the columns of $U_j \in \mathbb{M}_{N,j}$ and $W_j \in \mathbb{M}_j$ are the orthonormal left-singular and right-singular vectors of K_j, respectively, and $\Sigma_j \in \mathbb{M}_j$ with j singular values of K_j on its diagonal, respectively. Hence, one obtains that the orthogonal components $\boldsymbol{\theta}_j$ for every $1 \leq j \leq N$ have the norm

$$\begin{aligned} \|\boldsymbol{\theta}_{j+1}\| &= \|\,(\mathbb{I}_N - P_j^c)A^j\mathbf{C}\| = \|\,(\mathbb{I}_N - U_j\,\Sigma_j\,W_j^\top(W_j\Sigma_j U_j^\top U_j\,\Sigma_j\,W_j^\top)^{-1}W_j\Sigma_j U_j^\top)A^j\mathbf{C}\| \\ &= \|\,(\mathbb{I}_N - U_jU_j^\top)A^j\mathbf{C}\|. \end{aligned}$$

We call this singular value decomposition approach the ***orthogonal method***, as it explicitly removes dependence on the ill-conditioning that is now incorporated in the singular values matrices Σ_j.[3]

The results of our simulations with LESN models of size $N = 100$ are shown in Figure 3, which also include a rank estimation of K_N. As one can notice, even with a logarithmic ordinate axis, the decay for most random sampling distributions is faster than exponential when compared to the powers of the leading eigenvalue. Only when using a random orthogonal matrix the decay of $\|\boldsymbol{\theta}_j\|$ is close to $\rho(A)^{j-1}$, as shown in panel (d).

3. An alternative option is, of course, to use a standard linear projection argument i.e. least-squares to compute $\boldsymbol{\theta}_j$. Due to the Krylov structure, however, this method is very ill-conditioned.

Additionally, we make an important empirical observation: the value of $\|\boldsymbol{\theta}_j\|$ as a function of j is well approximated by the ordered cumulative product of the absolute values of eigenvalues of A. Our empirical finding can be seen in panels of Figure 3 by considering the dashed black line, which plots such cumulative eigenvalue product. This observation, combined with knowledge of the spectral properties of random matrices, allows getting a more precise understanding of the ill-conditioning of the reservoir autocovariance matrix, as we argue now.

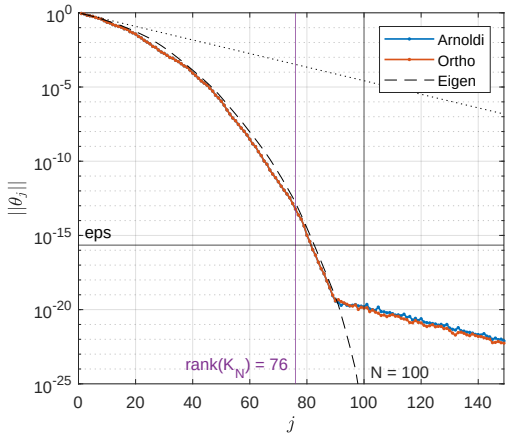
3.3.2 Random Matrix Theory Insights

A fundamental result of random matrix theory (RMT) is the celebrated *circular law*, which broadly speaking states that the (appropriately scaled) eigenvalues of families of random matrices are asymptotically uniformly distributed on the complex unit circle as $N \to \infty$ (see Tao 2012 for an introductory discussion). A general statement of the circular law for ensembles of matrices with i.i.d. entries with unit variance was given by Tao et al. (2010). Matrix ensembles with sparse entries also obey the circular law, as proven by Wood (2012) and Basak and Rudelson (2019). In particular, the degree of sparsity controls the probability of singularity, although in appropriate settings such probability remains exponentially small (Basak and Rudelson, 2017). Figure 6 in the Appendix shows the distribution of eigenvalues for commonly used ESN reservoir matrices. Notice that for ensembles of Gaussian, sparse Gaussian, and uniform entries, the associated eigenvalues have a close-to-uniform distribution on the complex unit circle. We highlight that attempts to use RMT to gain insights on reservoir models have already been made: Zhang et al. (2012) use the circular law to derive explicit bounds on spectral scaling factors; Couillet et al. (2016a) and Couillet et al. (2016b) apply results from random matrix theory to make performance analyses of linear echo state networks effected by exogenous noise. However, to the best of our knowledge, our empirical observations are new.

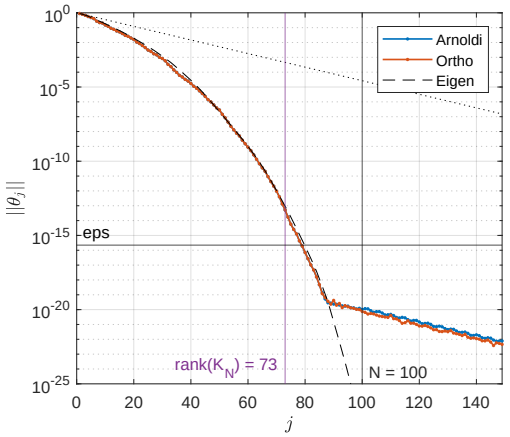
With the circular law in mind, in linear reservoirs where A is drawn randomly from standard matrix ensembles, we know that for its eigenvalues it approximately holds that $|\lambda_i|^2 \sim \mathcal{U}(0, \rho(A))$, $i \in \{1, \dots, N\}$. We can thus derive the following closed-form approximation, call it κ_j, for the value of $\|\boldsymbol{\theta}_j\|$, given by $\kappa_1 = 1$ and

$$\kappa_{j+1} = \sqrt{\rho(A)\,\frac{N!}{N^j(N-j)!}}\,, \quad \forall j \geq 1.$$

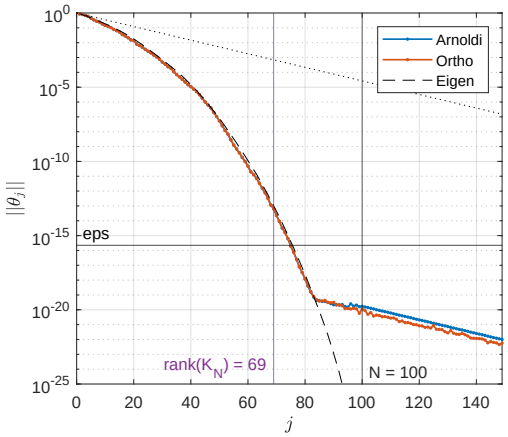
This expression can be derived easily by noting that $|\lambda_i|$ are approximately distributed as $\sqrt{\rho(A)Z_i}$ where $Z_i \sim \mathcal{U}(0,1)$. When N is large, we may further approximate realizations $(Z_i)_{i=1}^N$ with a uniform grid of knots over $(0,1)$. Computing the cumulative product of these knots in descending order gives the formula for κ_j above. It shows that the decay of $\|\boldsymbol{\theta}_j\|$ can be indeed much faster than that of powers of $\rho(A)$ under the circular law. In theory, a sharper formula could be derived by noting that most eigenvalues of a random matrix come in conjugate complex pairs, so knots should also be chosen in couples. This would require knowing the expected ratio of real to complex eigenvalues of a random matrix ensemble, which is beyond the scope of this discussion. Yet, as shown with the dashed black lines in Figure 3, we empirically find that our RMT approximation is remarkably precise at fitting the faster-than-exponential decay of $\|\boldsymbol{\theta}_j\|$.



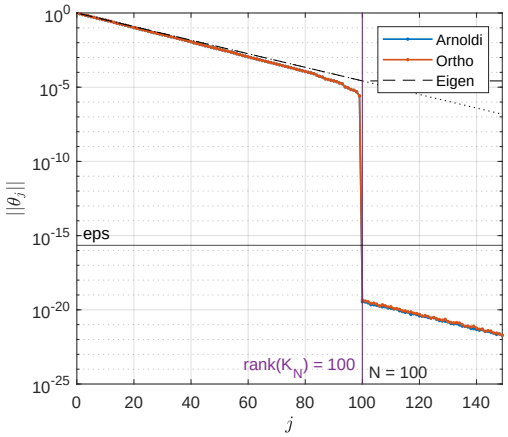
(a) $A_{ij} \sim$ i.i.d. $\mathcal{N}(0,1)$



(b) $A_{ij} \sim$ i.i.d. $\mathcal{U}(-1,1)$



(c) $A_{ij} \sim$ i.i.d. $sp\mathcal{N}(0,1,0.1)$



(d) $A \sim \mathcal{O}(\mathcal{N}(0,1))$

Figure 3: Krylov subspace squeezing effects as measured using the norm of the orthogonal component for reservoir matrix $A = (A_{ij}) \in \mathbb{M}_N$, $\rho(A) = 0.9$, sampled $\mathcal{N}(0,1)$ in (a), $\mathcal{U}(-1,1)$ in (b), sparse standard Gaussian with the degree of sparsity 0.1, $sp\mathcal{N}(0,1,0.1)$, in (c), and orthogonal standard Gaussian in (d), and for Krylov matrix $K_m \in \mathbb{M}_{N,m}$, where in all plots $N = 100$ and $m = 5N$. Input mask is $\mathbf{C} = \boldsymbol{\iota}_N = (1,\ldots,1)^\top \in \mathbb{R}^N$. The black *dotted* line shows the exponential decay of leading eigenvalue $\rho(A)$, while the black *dashed* line illustrates the approximate decay law derived using random matrix theory in Section 3.3.2. A solid black horizontal line shows the numerical double-precision of floating numbers in MATLAB, $eps = 2^{-52} \approx 2.22 \times 10^{-16}$.

3.4 Subspace Methods

Now that we have discussed the ill-posed nature of inverse problems involving $G_{\mathbf{x}}$ and how it relates to its Krylov structure, we propose two approaches that can correctly recover the full memory capacity as well as individual lag-τ capacities. Our methods boil down to the idea of using appropriate matrix decompositions to remove those parts of the singular values spectrum from normalized reservoir autocovariance $G_{\mathbf{x}}$ that lead to its ill-conditioning.

3.4.1 Orthogonalized Subspace Method

We start by recalling again the expression of the τ-lag memory capacity of the LESN given in (2.18), namely

$$\mathrm{MC}_\tau = \mathbf{C}^\top (A^\tau)^\top G_{\mathbf{x}}^{-1} A^\tau \mathbf{C}, \tag{3.6}$$

where, as we explained in Subsection 3.2, $G_{\mathbf{x}}$ can be approximated by

$$\widetilde{G}_{\mathbf{x}} = K_m K_m^\top \tag{3.7}$$

with $K_m \in \mathbb{M}_{N,m}$ a Krylov matrix with the column dimension truncated up to m as in (3.2). In this case, the approximate memory capacities in (3.6) can be computed as the diagonal elements of the following matrix:

$$P_m^r := K_m^\top (K_m K_m^\top)^{-1} K_m. \tag{3.8}$$

It is easy to see that this is a projection matrix corresponding to the projection operator $P_m^r : \mathbb{R}^m \longrightarrow \mathcal{C}(K_m^\top) \subset \mathbb{R}^m$. Using the singular value decomposition $K_m = U_m \, \Sigma_m \, W_m^\top$, with $U_m \in \mathbb{M}_N$ full-rank orthogonal with the left-singular vectors of K_m as columns, $\Sigma_m \in \mathbb{M}_N$ diagonal, and $W_m \in \mathbb{M}_{m,N}$ with the right-singular orthonormal vectors of K_m as columns (notice that this SVD is different from the one used in (3.5)), we write (3.8) as

$$P_m^r = W_m \Sigma_m U_m^\top (U_m \Sigma_m W_m^\top W_m \Sigma_m U_m^\top)^{-1} U_m \Sigma W_m^\top = W_m W_m^\top.$$

Therefore each τ-lag memory capacity of the LESN, $1 \leq \tau \leq m$, is well approximated by $(P_m^r)_{\tau,\tau}$.

It is important to underline that this method of memory capacity computation does not suffer from any of the previously mentioned matrix or linear system inversion issues, as it sidesteps the computation of $G_{\mathbf{x}}^{-1}$ altogether. The core idea is to explicitly exploit the subspace structure of the Krylov matrix K_m and, by using the singular value decomposition, to extract the projection matrix associated with the LESN memory capacity. We term this approach the ***orthogonalized subspace method*** (OSM) and define

$$\mathrm{MC}_\tau^{\mathrm{OSM}} = (W_m W_m^\top)_{\tau,\tau}. \tag{3.9}$$

Figures 4-5 show that the orthogonalized subspace method computes memory curves consistent with full memory. One also notices one of the downsides of this method when inspecting the memory curves recovered by OSM in all the panels of Figure 4 and in subfigure (a) and (b) of Figure 5. More precisely, OSM results in memory capacity curves that need not be monotonically decreasing. This is in contrast to the known monotonicity of memory proven in Jaeger (2002). A reason for this is that, while the subspace methods avoid a costly and

unstable matrix inversion, it still relies on the computation of singular value decomposition factors. Formula (3.9) does not guarantee that the diagonal entries are numerically non-increasing. Monotonicity hinges on recursively identifying the largest leading singular direction at each step of the decomposition. Accordingly, the accuracy of the estimated MC_τ is tied to the accuracy of the singular value decomposition, and the ill-conditioning of K_m still plays some role in it.

3.4.2 Averaged Orthogonalized Subspace Method

Finally, we propose an improved version of our subspace memory computation method that exploits the input mask memory neutrality result established in Proposition 6. Our goal is to leverage this property to produce a better approximation of MC_τ that is also monotonic.

We first notice that even though the expression of the true memory capacity MC_τ in (3.1) does not depend on $\mathbf{C}$, its numerical computation with OSM in (3.9) is impacted by the input mask. To recover the true memory capacity out of (3.9), one is ultimately interested in computing $\mathbb{E}_{\mathbf{C}}[(W_m W_m^\top)_{\tau,\tau}]$, $1 \leq \tau \leq m$, $m \in \mathbb{N}$, which, by Proposition 6, should not depend on a particular choice of the distribution $p_{\mathbf{C}}$ of the input mask. Although one can potentially choose any $p_{\mathbf{C}}$ that would allow one to evaluate this integral, we do not find obtaining its expression in a closed form feasible. Our proposal is to adhere to the sample estimator or the Monte Carlo estimator of $\mathbb{E}_{\mathbf{C}}[(W_m W_m^\top)_{\tau,\tau}]$, $1 \leq \tau \leq m$, $m \in \mathbb{N}$ which we will call the ***averaged orthogonalized subspace method***, or simply OSM+.

More explicitly, consider a sample of L independent and identically distributed according to some arbitrary chosen law $p_{\mathbf{C}}$ input masks $\{\mathbf{C}^{(1)}, \ldots, \mathbf{C}^{(L)}\}$, and using (3.9) construct the following memory capacity curve estimator:

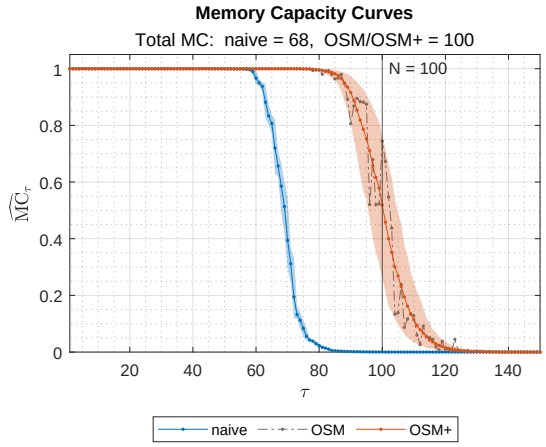
$$\mathrm{MC}_{L,\tau}^{\mathrm{OSM+}} = \frac{1}{L}\sum_{\ell=1}^{L}\left(W_m^{(\ell)} W_m^{(\ell)\top}\right)_{\tau,\tau}, \quad 1 \leq \tau \leq m, \tag{3.10}$$

for which the weak law of large numbers implies that

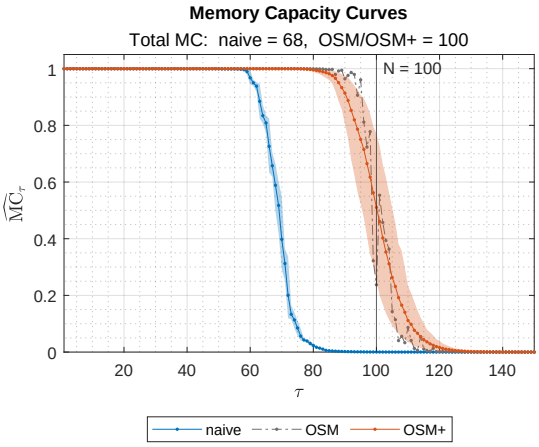
$$\mathrm{MC}_{L,\tau}^{\mathrm{OSM+}} \xrightarrow[L\to\infty]{p} \mathbb{E}_{\mathbf{C}}[(W_m W_m^\top)_{\tau,\tau}], \quad 1 \leq \tau \leq m.$$

As mentioned above, one of the key advantages of this construction is the fact that the OSM+ method allows choosing any type of $p_{\mathbf{C}}$ as long as the conditions of Proposition 6 are satisfied. Moreover, OSM+ is straightforward to implement numerically, as shown by the pseudo-code in Algorithm 1. Note that construction of the Krylov matrix can be done iteratively, and therefore the most computationally expensive operation is the singular value decomposition. Figures 4 and 5 show that the averaged subspace memory curves produced by OSM+ are indeed monotonic, in contrast to the one-step subspace approximation produced by OSM. Here, we do not make any suggestion as for the choice of the distribution for the entries of $\mathbf{C}$ since Figure 4 indicates that common choices yield very similar results for moderate resampling size $L = 1000$.

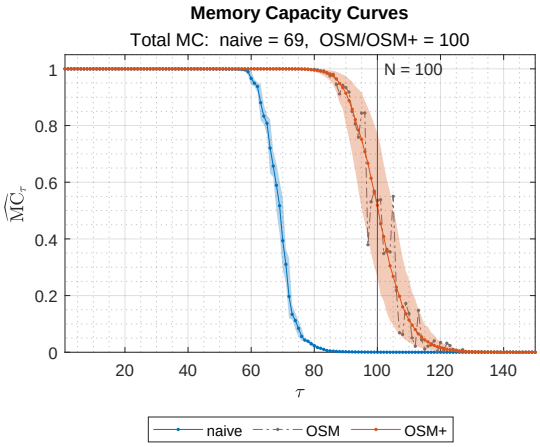
We emphasize that the naïve memory capacity computation discussed in Section 2.4 is able to recover full memory *only* for some particular choices of the connectivity architectures for which, by construction, the ill-conditioning problem is not pronounced. Indeed, the subplots (c) and (d) in Figure 5 indicate that all three methods, namely naïve, OSM, and OSM+, in these cases correctly quantify full memory of linear recurrent networks.



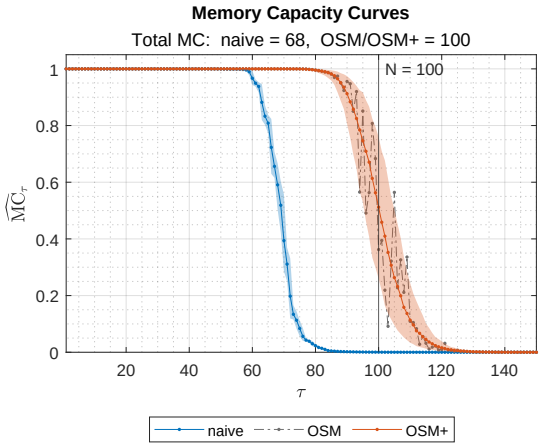
(a) $\mathbf{C} \sim$ i.i.d. $\mathcal{N}(0,1)$



(b) $\mathbf{C} \sim$ i.i.d. $\mathcal{U}(-1,1)$

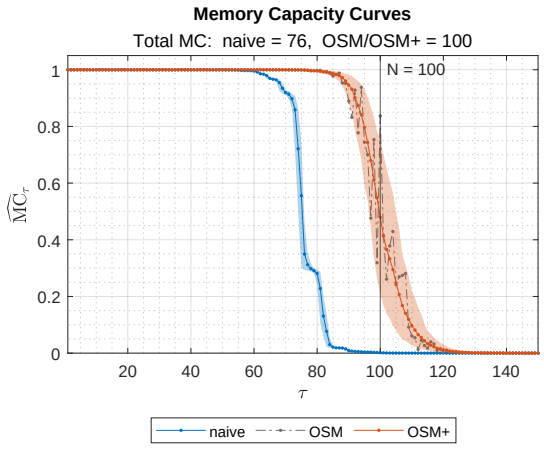


(c) $\mathbf{C} \sim$ i.i.d. $sp\mathcal{N}(0,1,0.1)$

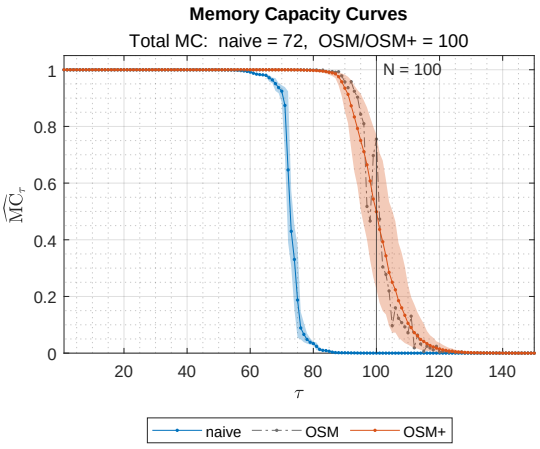


(d) $\mathbf{C} \sim$ i.i.d. $sp\mathcal{U}(0,1,0.1)$

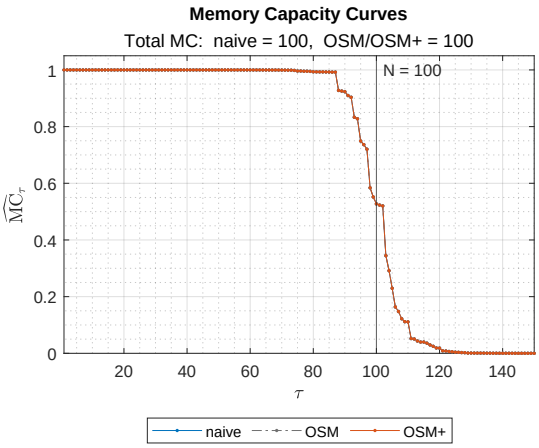
Figure 4: Memory capacity curves of LESNs with connectivity matrix $A = (A_{ij}) \in \mathbb{M}_N$ with $\rho(A) = 0.9$. In all panels $A_{i,j}$ are sampled as i.i.d. degree 0.1 sparse standard normal, $sp\mathcal{N}(0,1,0.1)$, and the input mask $\mathbf{C} = (c_i) \in \mathbb{R}^N$ is sampled as $\mathcal{N}(0,1)$ in (a), $\mathcal{U}(-1,1)$ in (b), degree 0.1 sparse Gaussian, $sp\mathcal{N}(0,1,0.1)$, in (c), and degree 0.1 sparse uniform, $sp\mathcal{U}(0,1,0.1)$, in (d). $\mathbf{C}$ is normalized after sampling to have a unit norm. Total MC is estimated as the sum of MC_τ's up to $1.5N$ terms. For OSM+ the input mask $\mathbf{C}$ is resampled $L = 1000$ times to compute the average memory curve (lines) and 90% frequency bands for MC_τ (shaded).



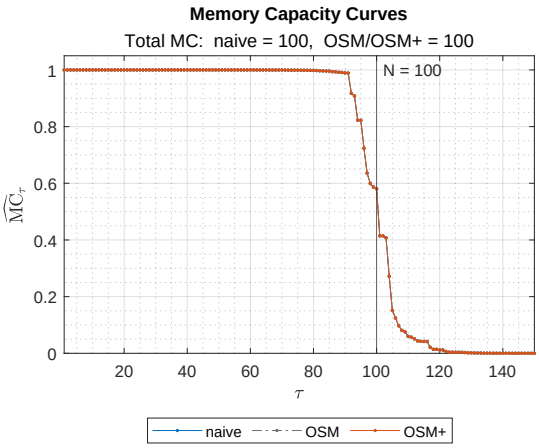
(a) $A \sim$ i.i.d. $\mathcal{N}(0,1)$



(b) $A \sim$ i.i.d. $\mathcal{U}(-1,1)$



(c) $A \sim \ \mathcal{O}(\mathcal{N}(0,1))$



(d) $A \sim \ sp_C\mathcal{N}(0,1,0.1,0.7)$

Figure 5: Memory capacity curves of LESNs with input mask $\mathbf{C} = (c_i) \in \mathbb{R}^N$ and connectivity matrix $A = (A_{ij}) \in \mathbb{M}_N$, $\rho(A) = 0.9$, sampled from different standard distributions (in panel (d) $sp_C\mathcal{N}(0,1,0.1,0.7)$ stands for sparse standard Gaussian with sparsity degree 0.1 and condition number 0.7). In all panels $c_i \sim$ i.i.d. $sp\mathcal{N}(0,1,0.1)$. $\mathbf{C}$ is normalized after sampling to have a unit norm. Total MC is computed as the sum of MC_τ's up to $1.5N$ terms. For OSM+ the input mask $\mathbf{C}$ is resampled $L = 1000$ times to compute the average memory curve (lines) and 90% frequency bands for MC_τ (shaded).

Algorithm 1: Averaged Orthogonalized Subspace Method (OSM+)

Input : Reservoir connectivity matrix $A \in \mathbb{M}_N$, distribution $p_{\mathbf{C}}$ of input matrix $\mathbf{C}$, Krylov matrix truncation order $m \in \mathbb{N}$, sampling budget $L \in \mathbb{N}$

Output: Memory capacity curve MC_τ for $0 \leq \tau \leq m$

```
mc_curve ← zeros(m+1, 1);                          # initialize MC curve vector
for ℓ ← 1 to L do
    C^(ℓ) ← sample_rand_matrix(p_C; N, 1);          # sample input matrix
    K_m^(ℓ) ← ( C^(ℓ) | A C^(ℓ) | A^2 C^(ℓ) | ... | A^(m-1) C^(ℓ) );   # construct Krylov matrix
    U_m^(ℓ), Σ_m^(ℓ), W_m^(ℓ)ᵀ ← svd(K_m^(ℓ));
    mc_curve ← mc_curve + L^(-1) diag(W_m^(ℓ) W_m^(ℓ)ᵀ);   # update estimate
```

4. Discussion

Given the fact that in this paper we have recalled (Section 2) and newly introduced (Section 3) many methods to compute the memory capacity of linear recurrent networks, specifically focusing on LESN models, we now wish to provide an overview of the key insights we have gathered by comparing them.

Simulations and naïve algebraic methods are both plagued by significant issues. In the former case, estimating moments of a stochastic process with simulated data always introduces some positive bias in the calculation of MC_τ, yielding memory curves that are inconsistent with the theoretical properties of memory. In the latter case, naïve algebraic applications of close-form formulas for memory eventually resort to inverting generally ill-poised covariance matrices, see (2.19)–(2.20) and Figure 2. The numerical instability in the inversion of $G_{\mathbf{x}}$ is the core issue with these approaches, and it is unavoidable by all techniques that directly rely on expression (2.7). Indeed, this means that simulation methods, too, are eventually impacted, as in extremely large simulations the conditioning of $\widehat{\mathrm{Var}}(\mathbf{x}_t)$ is close to that of $\Gamma_{\mathbf{x}}$, and thus, ultimately, $G_{\mathbf{x}}$.

While our OSM and, especially, OSM+ proposals are theoretically grounded and numerically well-conditioned approaches to estimating MC_τ, it is important also to mention that in this paper we do not provide theoretical results on the convergence properties of these algorithms. From a practical perspective, it would be interesting to derive a rate of convergence for subspace methods as $m \to \infty$ and τ is fixed or $\tau \to \infty$. The former seems easier, while the latter seems useful in providing a better understanding of how memory behaves at the "state dimension boundary", $\tau \approx N$, as $N \to \infty$, too. We leave these developments to future work.

Lastly, we mention how our results may be generalized to ***forecasting capacity***. Following again Gonon et al. (2020), recall that the forecasting capacity of an ESN is given by

$$\mathrm{FC}_\tau = \frac{\mathrm{Cov}(z_t, \mathbf{x}_{t-\tau})\Gamma_{\mathbf{x}}^{-1}\mathrm{Cov}(\mathbf{x}_{t-\tau}, z_t)}{\mathrm{Var}(z_t)}, \quad \tau \in \mathbb{N}_+, \tag{4.1}$$

where the states $\mathbf{x}_{t-\tau}$ are now lagged and not the inputs. Then, FC_τ is a linear measure of the predictability of future inputs with respect to the currently available state. Following

our discussion above on the implications of relying on (2.7), computations of FC_τ based on either simulations or naïve algebraic derivations are bound to provide inherently poor results. Although it is, in principle, possible to extend our OSM(+) methods also to compute FC_τ in a robust fashion, Corollary 3.5 in Gonon et al. (2020) proves that, for generic ESN models (not necessarily linear), if $(z_t)_{t\in\mathbb{Z}_-}$ is a sequence of independent random variables, then $\mathrm{FC}_\tau = 0$ and thus $\mathrm{FC} := \sum_{\tau=0}^{\infty} \mathrm{FC}_\tau = 0$. Therefore, for other types of stochastic inputs, forecasting capacity should be analyzed not as an inherent property of the (L)ESN model, but rather as a quantity based on the interaction between the model and inputs.

5. Conclusion

In this paper, we have provided an overview of the existing literature on memory capacity measures for recurrent neural networks and the approaches that have been extensively used in designing memory-optimal network architectures.

We have focused on explaining and providing solutions for what we call the linear memory gap, which refers to the difference between empirically measured memory capacities and their provable theoretical values. We have demonstrated that this discrepancy arises due to numerical artifacts that have been overlooked in previous studies.

We propose robust techniques for the accurate estimation of memory capacity, which result in full memory results for linear RNNs, as should be generically expected. Our findings suggest that previous efforts to optimize memory capacity for linear recurrent networks may have been plagued with numerical artifacts, leading to incorrect results. We base our findings on the fact that the capacities of linear systems are generically full, disregarding the particular choice of architecture. We also show that the memory capacity is neutral to the choice of the input mask. We propose two orthogonalized subspace methods that allow empirically recovering the full memory of linear systems and render results consistent with the theory.

We hope, with this conclusive work, to close the door to forthcoming attempts at memory optimization for linear RNNs that are not justified from a theoretical point of view.

Acknowledgments

GB and JPO thank the hospitality and the generosity of the University of St. Gallen, where part of this work was completed. LG thanks the Nanyang Technological University for the hospitality, which made some of this work possible. GB thanks Konstantin Usevich for the helpful discussion of algebraic insights regarding the empirical conjecture on ordered eigenvalue products. JPO acknowledges financial support from the Nanyang Technological University (grant number 020870-00001) and the Swiss National Science Foundation (grant number 200021_175801/1).

Appendix A. Memory Neutrality to Input Mask Under Stationarity

We now show that we can generalize Proposition 6 to the case of weakly stationary input processes that are not necessarily white noises. This provides a discrete-time counterpart to the result in Hermans and Schrauwen (2010) and allows us to apply our memory estimation methods in more general setups in which just the stationarity of the input is needed.

Theorem 9 *Under the controllability assumptions of Proposition 1, for any weakly stationary input $(z_t)_{t\in\mathbb{Z}}$ (not necessarily white noise), the memory of a linear echo state network is neutral to the choice of the input mask* $\mathbf{C}$*.*

Proof We shall mimic the proof of Proposition 6. We start by noticing that under the assumptions of the theorem, the stationarity of the input process implies stationarity of the associated states process $(\mathbf{x}_t)_{t\in\mathbb{Z}}$ as well as of the joint process $(z_{t+\tau}, \mathbf{x}_t)_{t\in\mathbb{Z}}$ for any $\tau \in \mathbb{N}$ (see Corollary 2.4 in Gonon et al. 2020), and calculate

$$\begin{aligned}\operatorname{Cov}(\mathbf{x}_t, z_{t+\tau}) = \operatorname{Cov}(\mathbf{x}_0, z_\tau) &= \mathbb{E}\left[\sum_{j=0}^{\infty} A^j \mathbf{C} z_{-j} z_\tau\right] = \sum_{k=1}^{N}\sum_{j=0}^{\infty} \lambda_k^j \mathbb{E}\left[z_{-j} z_\tau\right] c_k \mathbf{v}_k \\ &= \sum_{k=1}^{N}\left(\sum_{j=0}^{\infty} \lambda_k^j \gamma(\tau - j)\right) c_k \mathbf{v}_k = \sum_{k=1}^{N} g_k(\tau) c_k \mathbf{v}_k,\end{aligned}$$

The state autocovariance matrix is given by

$$\begin{aligned}\Gamma_{\mathbf{x}} &= \mathbb{E}\left[\left(\sum_{i=0}^{\infty} A^i \mathbf{C} z_{-i}\right)\left(\sum_{j=0}^{\infty} A^j \mathbf{C} z_{-j}\right)^{\top}\right] \\ &= \mathbb{E}\left[\sum_{i=0}^{\infty} A^i \mathbf{C}\mathbf{C}^{\top}(A^{\top})^i z_{-i}^2 + \sum_{j\geq 1}\sum_{i=0}^{\infty}\left\{A^i \mathbf{C}\mathbf{C}^{\top}(A^{\top})^{i+j} + A^{i+j}\mathbf{C}\mathbf{C}^{\top}(A^{\top})^i\right\} z_{-i} z_{-i-j}\right] \\ &= \sum_{i=0}^{\infty} A^i \mathbf{C}\mathbf{C}^{\top}(A^{\top})^i \gamma(0) + \sum_{j\geq 1}\sum_{i=0}^{\infty}\left\{A^i \mathbf{C}\mathbf{C}^{\top}(A^{\top})^{i+j} + A^{i+j}\mathbf{C}\mathbf{C}^{\top}(A^{\top})^i\right\}\gamma(j).\end{aligned}$$

We now analyze all three summands separately:

$$\begin{aligned}\sum_{i=0}^{\infty} A^i \mathbf{C}\mathbf{C}^{\top}(A^{\top})^i \gamma(0) &= \gamma(0)\sum_{k,l=1}^{N} c_k \overline{c}_l \frac{1}{1-\lambda_k\overline{\lambda}_l}\mathbf{v}_k \mathbf{v}_l^*, \\ \sum_{i=0}^{\infty} A^i \mathbf{C}\mathbf{C}^{\top}(A^{\top})^{i+j} \gamma(j) &= \gamma(j)\sum_{k,l=1}^{N} c_k \overline{c}_l \frac{\overline{\lambda}_l^j}{1-\lambda_k\overline{\lambda}_l}\mathbf{v}_k \mathbf{v}_l^*, \\ \sum_{i=0}^{\infty} A^{i+j} \mathbf{C}\mathbf{C}^{\top}(A^{\top})^i \gamma(j) &= \gamma(j)\sum_{k,l=1}^{N} c_k \overline{c}_l \frac{\lambda_k^i}{1-\lambda_k\overline{\lambda}_l}\mathbf{v}_k \mathbf{v}_l^*,\end{aligned}$$

and can simplify $\Gamma_{\mathbf{x}}$ as

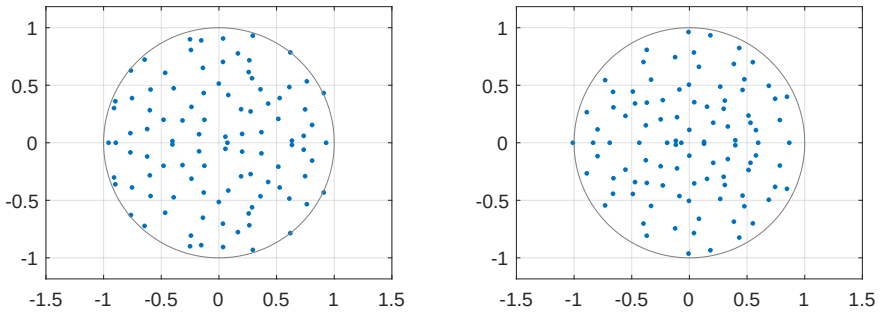
$$\Gamma_{\mathbf{x}} = \sum_{k,l=1}^{N} c_k \bar{c}_l \left(\frac{\gamma(0)}{1 - \lambda_k \overline{\lambda}_l} + \sum_{j \geq 1} \gamma(j) \frac{\lambda_k^j + \overline{\lambda}_l^j}{1 - \lambda_k \overline{\lambda}_l} \right) \mathbf{v}_k \mathbf{v}_l^* = \sum_{k,l=1}^{N} c_k \bar{c}_l \frac{h_{k,l}}{1 - \lambda_k \overline{\lambda}_l} \mathbf{v}_k \mathbf{v}_l^*.$$

We can now mimic the derivation in the proof of Proposition 6 as follows

$$\begin{aligned}
\mathrm{MC}_\tau &= \gamma(0)^{-1} \mathrm{Cov}(\mathbf{x}_t, z_{t-\tau})^* \Gamma_{\mathbf{x}}^{-1} \mathrm{Cov}(\mathbf{x}_t, z_{t-\tau}) \\
&= \gamma(0)^{-1} \left(\sum_{k=1}^{N} g_k(\tau) c_k \mathbf{v}_k \right)^* \left(\sum_{k,l=1}^{N} c_k \bar{c}_l \frac{h_{k,l}}{1 - \lambda_k \overline{\lambda}_l} \mathbf{v}_k \mathbf{v}_l^* \right)^{-1} \left(\sum_{k=1}^{N} g_k(\tau) c_k \mathbf{v}_k \right) \\
&= \gamma(0)^{-1} \mathbf{c}^* G(\tau)^* \left(\mathrm{diag}\left(\mathbf{c}\right) \left(\frac{h_{k,l}}{1 - \lambda_k \overline{\lambda}_l} \right)_{k,l}^{N} \mathrm{diag}\left(\mathbf{c}^*\right) \right)^{-1} G(\tau)\, \mathbf{c} \\
&= \boldsymbol{\iota}_N^\top G(\tau)^* \left(\gamma(0) \left(\frac{h_{k,l}}{1 - \lambda_k \overline{\lambda}_l} \right)_{k,l}^{N} \right)^{-1} G(\tau)\, \boldsymbol{\iota}_N.
\end{aligned}$$

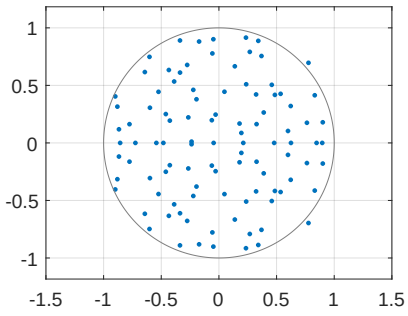
Notice that the final expression does not depend on $\mathbf{C}$, which proves the neutrality of the memory capacity with respect to the choice of the input mask, as required. ■

Appendix B. Additional Plots

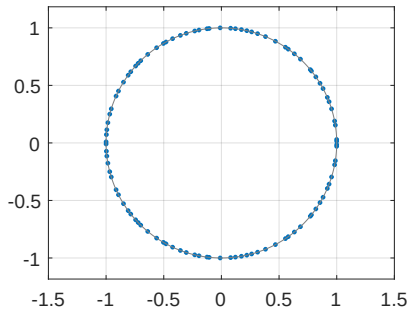


(a) $A_{ij} \sim$ i.i.d. $\mathcal{N}(0,1)$

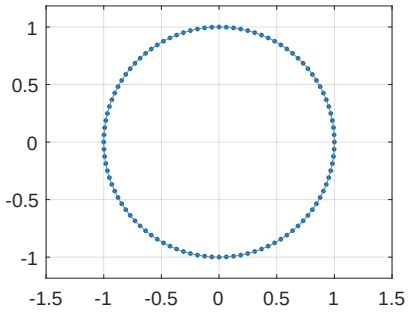
(b) $A_{ij} \sim$ i.i.d. $sp\mathcal{N}(0,1,0.1)$



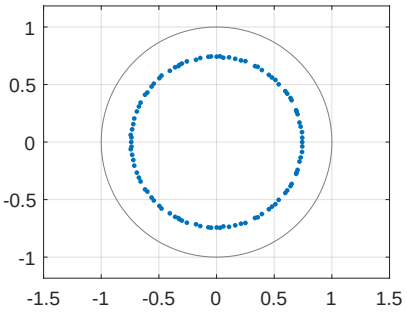
(c) $A_{ij} \sim$ i.i.d. $\mathcal{U}(-1,1)$



(d) $A_{ij} \sim \mathcal{O}(\mathcal{N}(0,1))$



(e) $A_{ij} \sim \text{cyclic}(N)$



(f) $A_{ij} \sim$ i.i.d. $sp_C\mathcal{N}(0,1,0.1,0.7)$

Figure 6: Eigenvalues (blue) for random and non-random reservoir matrices and the complex unit circle (gray), $N = 100$. For specifications with entries $A_{ij} \sim$ i.i.d. $\mathcal{N}$, $sp\mathcal{N}$ and $\mathcal{U}$ (upper row) the matrices are normalized according to the circular law rates $N^{-1/2}$, $(0.1N)^{-1/2}$ and $(N/3)^{-1/2}$, respectively. In (f) $sp_C\mathcal{N}(0,1,0.1,0.7)$ stands for sparse standard Gaussian with sparsity degree 0.1 and condition number 0.7.

References

B. Acciaio, A. Kratsios, and G. Pammer. Designing universal causal deep learning models: The geometric (hyper)transformer. *Mathematical Finance*, 34:671–735, 2024.

P. V. Aceituno, Y. Gang, and Y.-Y. Liu. Tailoring Echo State Networks for optimal learning. *iScience*, 23(9), 2020.

W. E. Arnoldi. The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quarterly of Applied Mathematics*, 9(1):17–29, 1951.

P. Barancok and I. Farkas. Memory capacity of input-driven echo state networks at the edge of chaos. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, pages 41–48, 2014.

A. Basak and M. Rudelson. Invertibility of sparse non-Hermitian matrices. *Advances in Mathematics*, 310:426–483, 2017.

A. Basak and M. Rudelson. The circular law for sparse non-Hermitian matrices. *The Annals of Probability*, 4(47):2359–2416, 2019.

M. Bellalij, G. Meurant, and H. Sadok. The distance of an eigenvector to a Krylov subspace and the convergence of the Arnoldi method for eigenvalue problems. *Linear Algebra and its Applications*, 504:387–405, 2016.

P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer-Verlag, 2006.

A. Charles, H. Yap, and C. Rozell. Short term network memory capacity via the restricted isometry property. *Neural Computation*, 26, 2014.

A. S. Charles, D. Yin, and C. J. Rozell. Distributed sequence memory of multidimensional inputs in recurrent networks. Technical report, 2017.

X. Chen, M. Xu, and W. B. Wu. Covariance and precision matrix estimation for high-dimensional time series. *The Annals of Statistics*, 41(6):2994–3021, 2013.

R. Couillet, G. Wainrib, H. T. Ali, and H. Sevi. A random matrix approach to echo-state neural networks. *Proceedings of The 33rd International Conference on Machine Learning*, pages 517–525, 2016a.

R. Couillet, G. Wainrib, H. Sevi, and H. T. Ali. The asymptotic performance of linear echo state neural networks. *Journal of Machine Learning Research*, 17(178):1–35, 2016b.

J. Dambre, D. Verstraeten, B. Schrauwen, and S. Massar. Information processing capacity of dynamical systems. *Scientific reports*, 2(514), 2012.

M. Eiermann and O. G. Ernst. Geometric aspects of the theory of Krylov subspace methods. *Acta Numerica*, pages 251–312, 2001.

I. Farkas, R. Bosak, and P. Gergel. Computational analysis of memory capacity in echo state networks. *Neural Networks*, 83:109–120, 2016.

L. Galimberti, G. Livieri, and A. Kratsios. Designing universal causal deep learning models: the case of infinite-dimensional dynamical systems from stochastic analysis. *arXiv preprint arXiv:2210.13300*, 2022.

C. Gallicchio. Short-term memory of Deep RNN. *arXiv preprint arXiv:1802.00748v1*, 2018.

C. Gallicchio. Sparsity in reservoir computing neural networks. *arXiv preprint arXiv:2006.02957v1*, 2020.

C. Gallicchio, A. Micheli, and L. Pedrelli. Deep reservoir computing: a critical experimental analysis. *Neurocomputing*, (April):87–99, 2017.

S. Ganguli, D. Huh, and H. Sompolinsky. Memory traces in dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, 105(48):18970–5, 2008.

L. Gonon and J.-P. Ortega. Reservoir computing universality with stochastic inputs. *IEEE Transactions on Neural Networks and Learning Systems*, 31(1):100–112, 2020.

L. Gonon and J.-P. Ortega. Fading memory echo state networks are universal. *Neural Networks*, 138:10–13, 2021.

L. Gonon, L. Grigoryeva, and J.-P. Ortega. Memory and forecasting capacities of nonlinear recurrent networks. *Physica D*, 414(132721):1–13, 2020.

A. Goudarzi, S. Marzen, P. Banda, G. Feldman, M. R. Lakin, C. Teuscher, and D. Stefanovic. Memory and information processing in recurrent neural networks. Technical report, Portland State University, 2016.

L. Grigoryeva and J.-P. Ortega. Echo state networks are universal. *Neural Networks*, 108: 495–508, 2018.

L. Grigoryeva and J.-P. Ortega. Dimension reduction in recurrent networks by canonicalization. *Journal of Geometric Mechanics*, 13(4):647–677, 2021.

L. Grigoryeva, J. Henriques, L. Larger, and J.-P. Ortega. Stochastic time series forecasting using time-delay reservoir computers: performance and universality. *Neural Networks*, 55:59–71, 2014.

L. Grigoryeva, J. Henriques, L. Larger, and J.-P. Ortega. Optimal nonlinear information processing capacity in delay-based reservoir computers. *Scientific Reports*, 5(12858):1–11, 2015.

L. Grigoryeva, J. Henriques, L. Larger, and J.-P. Ortega. Nonlinear memory capacity of parallel time-delay reservoir computers in the processing of multidimensional signals. *Neural Computation*, 28:1411–1451, 2016a.

L. Grigoryeva, J. Henriques, and J.-P. Ortega. Reservoir computing: information processing of stationary signals. In *Proceedings of the 19th IEEE International Conference on Computational Science and Engineering*, pages 496–503, 2016b.

L. Grigoryeva, A. G. Hart, and J.-P. Ortega. Learning strange attractors with reservoir systems. *Nonlinearity*, 36:4674–4708, 2023.

J. D. Hamilton. *Time series analysis*. Princeton University Press, Princeton, NJ, 1994.

Y. Hashimoto, I. Ishikawa, M. Ikeda, Y. Matsuo, and Y. Kawahara. Krylov subspace method for nonlinear dynamical systems with random noise. *Journal of Machine Learning Research*, 21:1–29, 2020.

D. Haviv, A. Rivkind, and O. Barak. Understanding and controlling memory in recurrent neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2663–2671, 2019.

M. Hermans and B. Schrauwen. Memory in linear recurrent neural networks in continuous time. *Neural Networks*, 23(3):341–55, 2010.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, second edition, 2013.

H. Jaeger. Short term memory in echo state networks. *Fraunhofer Institute for Autonomous Intelligent Systems. Technical Report*, 152, 2002.

H. Jaeger and H. Haas. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304(5667):78–80, 2004.

R. Kalman. Lectures on Controllability and Observability. In *Controllability and Observability*, pages 1–149. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

E. Koyuncu. Memorization capacity of neural networks with conditional computation. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

P. Lax. *Functional Analysis*. Wiley-Interscience, 2002.

Z. Li, J. Han, W. E, and Q. Li. On the curse of memory in recurrent neural networks: Approximation and optimization analysis. In *ICLR*, pages 1–43, 2021.

L. Livi, F. M. Bianchi, and C. Alippi. Determination of the edge of criticality in echo state networks through Fisher information maximization. 2016.

H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer-Verlag, Berlin, 2 edition, 2005.

G. Manjunath and J.-P. Ortega. Transport in reservoir computing. *Physica D: Nonlinear Phenomena*, 449:133744, 2023.

S. Marzen. Difference between memory and prediction in linear recurrent networks. *Physical Review E*, 96(3):1–7, 2017.

M. Matthews and G. Moschytz. The identification of nonlinear discrete-time fading-memory systems using neural network models. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 41(11):740–751, 1994.

M. B. Matthews. *On the Uniform Approximation of Nonlinear Discrete-Time Fading-Memory Systems Using Neural Network Models*. PhD thesis, ETH Zürich, 1992.

G. Meurant and J. Duintjer Tebbens. *Krylov Methods for Nonsymmetric Linear Systems*. Springer International Publishing, 2020.

S. Ortín and L. Pesquera. Tackling the trade-off between information processing capacity and rate in delay-based reservoir computers. *Frontiers in Physics*, 7:210, 2019.

S. Ortín and L. Pesquera. Delay-based reservoir computing: tackling performance degradation due to system response time. *Optics Letters*, 45(4):905–908, 2020.

S. Ortin, L. Pesquera, and J. M. Gutiérrez. Memory and nonlinear mapping in reservoir computing with two uncoupled nonlinear delay nodes. In T. Gilbert, M. Kirkilionis, and G. Nicolis, editors, *Proceedings of the European Conference on Complex Systems*, pages 895–899. Springer International Publishing Switzerland, 2012.

A. Rodan and P. Tino. Minimum complexity echo state network. *IEEE Transactions on Neural Networks*, 22(1):131–44, 2011.

A. Rodan and P. Tino. Simple deterministically constructed cycle reservoirs with regular jumps. *Neural Computation*, 24(7):1822–1852, 2012.

E. Sontag. *Mathematical Control Theory: Deterministic Finite Dimensional Systems*. Springer-Verlag, 1998.

E. D. Sontag. Kalman's controllability rank condition: from linear to nonlinear. In A. C. Antoulas, editor, *Mathematical System Theory*, pages 453–462. Springer, 1991.

T. Strauss, W. Wustlich, and R. Labahn. Design strategies for weight matrices of echo state networks. *Neural Computation*, 24(12):3246–3276, 2012.

I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *NeurIPS*, pages 3104–3112, 2014.

T. Tao. *Topics in Random Matrix Theory*. American Mathematical Society, 2012.

T. Tao, V. Vu, and M. Krishnapur. Random matrices: Universality of ESDs and the circular law. *The Annals of Probability*, 38(5):2023–2065, 2010.

P. Tino. Asymptotic Fisher memory of randomized linear symmetric echo state networks. *Neurocomputing*, 298:4–8, 2018.

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

E. E. Tyrtyshnikov. How bad are Hankel matrices? *Numerische Mathematik*, 67(2):261–269, 1994.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, pages 1–11, 2017.

R. Vershynin. Memory capacity of neural networks with threshold and rectified linear unit activations. *SIAM Journal on Mathematics of Data Science*, 2(4), 2020.

P. Verzelli, C. Alippi, and L. Livi. Echo State Networks with self-normalizing activations on the hyper-sphere. *Scientific Reports*, 9(1):13887, 2019.

P. Verzelli, C. Alippi, L. Livi, and P. Tino. Input-to-state representation in linear reservoirs dynamics. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12, 2021.

O. White, D. Lee, and H. Sompolinsky. Short-term memory in orthogonal neural networks. *Physical Review Letters*, 92(14):148102, 2004.

B. Whiteaker and P. Gerstoft. Reducing echo state network size with controllability matrices. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 32(7):73116, 2022a.

B. Whiteaker and P. Gerstoft. Memory in Echo State Networks and the controllability matrix rank. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3948–3952, 2022b.

P. M. Wood. Universality and the circular law for sparse random matrices. *The Annals of Probability*, 22(3):1266–1300, 2012.

F. Xue, Q. Li, and X. Li. The combination of circle topology and leaky integrator neurons remarkably improves the performance of echo state network on time series prediction. *PloS one*, 12(7):e0181816, 2017.

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, 2012.

D. Zhang and W. B. Wu. Gaussian approximation for high dimensional time series. *The Annals of Statistics*, 45(5):1895–1919, 2017.