Physica D 414 (2020) 132721

Contents lists available at ScienceDirect

Physica D

journal homepage: www.elsevier.com/locate/physd

Memory and forecasting capacities of nonlinear recurrent networks

Lukas Gonon [a,1], Lyudmila Grigoryeva [b,1], Juan-Pablo Ortega [c,d,*,1]

[a] Ludwig-Maximilians-Universität München, Faculty of Mathematics, Informatics and Statistics, Theresienstrasse 39, D-80333 Munich, Germany
[b] Universität Konstanz, Department of Mathematics and Statistics, Universitätstrasse 10, D-78457 Konstanz, Germany
[c] Universität Sankt Gallen, Faculty of Mathematics and Statistics, Bodanstrasse 6, CH-9000 Sankt Gallen, Switzerland
[d] Centre National de la Recherche Scientifique (CNRS), France

ARTICLE INFO

Article history:
Received 22 April 2020
Received in revised form 25 July 2020
Accepted 31 August 2020
Available online 5 September 2020
Communicated by B. Hamzi

Keywords:
Memory capacity
Forecasting capacity
Recurrent neural network
Reservoir computing
Echo state network (ESN)
Machine learning

ABSTRACT

The notion of memory capacity, originally introduced for echo state and linear networks with independent inputs, is generalized to nonlinear recurrent networks with stationary but dependent inputs. The presence of dependence in the inputs makes natural the introduction of the network forecasting capacity, that measures the possibility of forecasting time series values using network states. Generic bounds for memory and forecasting capacities are formulated in terms of the number of neurons of the nonlinear recurrent network and the autocovariance function or the spectral density of the input. These bounds generalize well-known estimates in the literature to a dependent inputs setup. Finally, for the particular case of linear recurrent networks with independent inputs it is proved that the memory capacity is given by the rank of the associated controllability matrix, a fact that has been for a long time assumed to be true without proof by the community.

1. Introduction

Memory capacities have been introduced in [1] in the context of recurrent neural networks in general and of echo state networks (ESNs) [2–4] in particular, as a way to quantify the amount of information contained in the states of a state-space system in relation with past inputs and as a measure of the ability of the network to retain the dynamic features of processed signals.

In the original definition, the memory capacity was defined as the sum of the coefficients of determination of the different linear regressions that use the state of the system at a given time as covariates and the values of the input at a given lagged time in the past as dependent variables. This notion has been the subject of much research in the reservoir computing literature [5–14] where most of the efforts have been concentrated in linear and echo state systems. An analytical expression of the capacity of time-delay reservoirs has been formulated in [15,16] and various proposals for optimized reservoir architectures can be obtained by maximizing the capacity as a function of reservoir hyperparameters [17–20]. Additionally, memory capacities have been extensively compared with other related concepts like Fisher information-based criteria [21–23].

All the above-mentioned works consider exclusively independent or white noise input signals and it is, to our knowledge, only in [24] (using autocorrelated inputs via a sparsity model), [25] (under strong high-order stationarity assumptions), [26] (using autocorrelated inputs via low-rank multi-input models), and [27] (for linear recurrent networks and inputs coming from a countable hidden Markov model), that the case involving dependent inputs has been treated. Since most signals that one encounters in applications exhibit some sort of temporal dependence, studying this case in detail is of obvious practical importance. Moreover, the presence of dependence makes pertinent considering not only memory capacities, but also the possibility to forecast time series values using network states, that is, ***forecasting capacities***. This notion has been introduced for the first time in [27] and studied in detail for linear recurrent networks under the hypothesis that the inputs are realizations of a countable hidden Markov model. That work shows, in particular, that linear networks optimized for memory capacity do not necessarily have a good forecasting capacity and vice versa.

This paper contains two main contributions. First, we extend the results on memory and forecasting capacities available in the literature exclusively for either linear recurrent or echo state networks and for uncorrelated/independent inputs *to general nonlinear systems and to dependent inputs that are only assumed to be stationary*. We show that under particular assumptions our bounds reduce to those well-known in the literature. More specifically, it is known since [1] that the memory capacity of an ESN or a linear recurrent network defined using independent

* Corresponding author at: Universität Sankt Gallen, Faculty of Mathematics and Statistics, Bodanstrasse 6, CH-9000 Sankt Gallen, Switzerland.
E-mail addresses: gonon@math.lmu.de (L. Gonon), Lyudmila.Grigoryeva@uni-konstanz.de (L. Grigoryeva), Juan-Pablo.Ortega@unisg.ch (J.-P. Ortega).
[1] All authors have contributed equally in the construction of this paper.

https://doi.org/10.1016/j.physd.2020.132721

inputs (we call this the *classical linear case*) is bounded above by the number of its output neurons or, equivalently, by the dimensionality of the corresponding state-space representation. We show that these memory capacity bounds in [1] immediately follow from our results. Second, in the linear case we reveal new relations between the memory capacity, spectral properties of the connectivity matrix of the network, and the so-called Kalman's characterization of the controllability of a linear system. More explicitly, [1] shows that the memory capacity is maximal if and only if Kalman's controllability rank condition [28–30] is satisfied. In this work we make a step further and prove that *in the classical case the memory capacity is given exactly by the rank of the controllability matrix*.

The paper is organized as follows:

- Section 2 introduces recurrent neural networks with linear readouts in relation with state-space representations. We focus on a large class of state-space systems that satisfy the so-called echo state property (ESP) and which guarantees that they uniquely determine an input/output system (also referred as *filter* in this paper). We recall well-known sufficient conditions for this property to hold, the notions of system morphism and isomorphism, and discuss how the non-uniqueness of state-space representations can be handled. We also carefully introduce the stationarity hypotheses that are invoked in the rest of the paper. Finally, in Proposition 2.5 we introduce an important technical result that shows that if we have a state system and an input for which the output process is covariance stationary and the corresponding covariance matrix is non-singular, then an isomorphic system representation exists whose corresponding state process is standardized, that is, the states have mean zero and covariance matrix equal to the identity. This standardization leads to systems that are easier to handle in terms of the computation of memory and forecasting capacities, which is profusely exploited later on in the main results of the paper.
- Section 3 contains the first main contribution of the paper. We first provide the definitions of the memory and forecasting capacities of nonlinear systems with linear readouts in the presence of stationary inputs and outputs. Second, Lemma 3.3 shows that the memory and forecasting capacities of state-space systems with linear readouts are invariant with respect to linear system morphisms; this is a technical tool used later on in Section 4. Finally, Theorem 3.4 provides bounds for the memory and the forecasting capacities of generic recurrent networks with linear readouts and with stationary inputs in terms of the dimensionality of the corresponding state-space representation and the autocovariance or the spectral density function of the input, which is assumed to be second-order stationary. These bounds reduce to those in [1] when the inputs are independent.
- Section 4 is exclusively devoted to the linear case. We study separately the cases in which the (time-independent) covariance matrices of the state process are invertible and non-invertible. In the regular case, explicit expressions for the memory and the forecasting capacities can be stated (see Proposition 4.1) in terms of the matrix parameters of the network (the so-called connectivity and input matrices) and the autocorrelation properties of the input. Moreover, in the *classical* case (linear recurrent network with independent inputs), these expressions yield interesting relations (see Proposition 4.3) between maximum memory capacity, spectral properties of the connectivity matrix of the network, and the so-called Kalman's characterization of the controllability of a linear state-space system [28]. This last condition has been already mentioned in [1] in relation with maximal capacity. When the state covariance matrix is singular, a completely different strategy is adopted based on using the invariance of capacities under linear system morphisms. Theorem 4.4 proves that *the memory capacity of a linear recurrent network with independent inputs is given by the rank of its controllability matrix*. This statement obviously generalizes the one established in [1]. Even though, to our knowledge, this is the first rigorous proof of the relation between network memory and the rank of the controllability matrix, that link has been for a long time part of the reservoir computing folklore. In particular, recent contributions are dedicated to the design of ingenuous configurations that maximize that rank [31–33].
- Section 5 concludes the paper and all the proofs are contained in the Appendix.

2. Recurrent neural networks with stationary inputs

The results in this paper apply to recurrent neural networks determined by state-space equations of the form:

$$\begin{cases}\mathbf{x}_t = F(\mathbf{x}_{t-1}, \mathbf{z}_t), & \text{(a)}\\ \mathbf{y}_t = h(\mathbf{x}_t) := \mathbf{W}^\top \mathbf{x}_t + \mathbf{a}, & \text{(b)}\end{cases} \tag{2.1}$$

for any $t \in \mathbb{Z}$. These two relations form a ***state-space system***, where the map $F : D_N \subset \mathbb{R}^N \times D_d \subset \mathbb{R}^d \longrightarrow D_N \subset \mathbb{R}^N$, $N, d \in \mathbb{N}$, is called the ***state map*** and $h : \mathbb{R}^N \longrightarrow \mathbb{R}^m$ the ***readout*** or ***observation*** map that, all along this paper, will be assumed to be affine, that is, it is determined just by a matrix $\mathbf{W} \in \mathbb{M}_{N,m}$ and a vector $\mathbf{a} \in \mathbb{R}^m$, $m \in \mathbb{N}$. The ***inputs*** $\{\mathbf{z}_t\}_{t\in\mathbb{Z}}$ of the system, with $\mathbf{z}_t \in D_d$, will be in most cases infinite paths of a discrete-time stochastic process. We note that, unlike what we do in this paper, the term ***recurrent neural network*** is used sometimes in the literature to refer exclusively to state-space systems where the state map F in (2.1a) is neural network-like, that is, it is the composition of a nonlinear activation function with an affine function of the states and the input.

We shall focus on state-space systems of the type (2.1a)–(2.1b) that determine an ***input/output*** system. This happens in the presence of the so-called ***echo state property (ESP)***, that is, when for any $\mathbf{z} \in (D_d)^{\mathbb{Z}}$ there exists a unique $\mathbf{y} \in (\mathbb{R}^m)^{\mathbb{Z}}$ such that (2.1a)–(2.1b) hold. In that case, we talk about the ***state-space filter*** $U_h^F : (D_d)^{\mathbb{Z}} \longrightarrow (\mathbb{R}^m)^{\mathbb{Z}}$ associated to the state-space system (2.1a)–(2.1b) defined by:

$$U_h^F(\mathbf{z}) := \mathbf{y},$$

where $\mathbf{z} \in (D_d)^{\mathbb{Z}}$ and $\mathbf{y} \in (\mathbb{R}^m)^{\mathbb{Z}}$ are linked by (2.1a) via the ESP. If the ESP holds at the level of the state equation (2.1a), we can define a ***state filter*** $U^F : (D_d)^{\mathbb{Z}} \longrightarrow (D_N)^{\mathbb{Z}}$ and, in that case, we have that

$$U_h^F := h \circ U^F.$$

It is easy to show that state and state-space filters are automatically causal and time-invariant (see [34, Proposition 2.1]) and hence it suffices to work with their restriction $U_h^F : (D_d)^{\mathbb{Z}_-} \longrightarrow (\mathbb{R}^m)^{\mathbb{Z}_-}$ to semi-infinite inputs and outputs. Moreover, U_h^F determines a ***state-space functional*** $H_h^F : (D_d)^{\mathbb{Z}} \longrightarrow \mathbb{R}^m$ as $H_h^F(\mathbf{z}) := U_h^F(\mathbf{z})_0$, for all $\mathbf{z} \in (D_d)^{\mathbb{Z}_-}$ (the same applies to U^F and H^F when the ESP holds at the level of the state equation). In the sequel we use the symbol $\mathbb{Z}_-$ to denote the negative integers including zero and $\mathbb{Z}^-$ without zero.

The echo state property has received much attention in the context of the so-called ***echo state networks (ESNs)*** [2–4,35,36] (see, for instance, [37–43]). Sufficient conditions for the ESP to hold in general systems have been formulated in [34,44–46], in most cases assuming that the state map is a contraction in the state variable. We now recall a result (see [45, Theorem 12] and [46, Proposition 1]) that ensures the ESP as well as a continuity property of the state filter which are important for the sequel. All along this paper and whenever in the presence of Cartesian products (finite or infinite) of topological spaces, continuity will be considered with respect to the product topology, that is, the coarsest topology that makes continuous all the canonical projections onto the individual factors (see [47, Chapter 2] for details).

Proposition 2.1. *Let $F : D_N \times D_d \longrightarrow D_N$ be a continuous state map such that D_N is a compact subset of $\mathbb{R}^N$ and F is a contraction on the first entry with constant $0 < c < 1$, that is,*

$$\|F(\mathbf{x}_1, \mathbf{z}) - F(\mathbf{x}_2, \mathbf{z})\| \leq c\|\mathbf{x}_1 - \mathbf{x}_2\|,$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in D_N$, $\mathbf{z} \in D_d$. Then, the associated system has the echo state property for any input in $(D_d)^{\mathbb{Z}_-}$. The associated filter $U^F : (D_d)^{\mathbb{Z}_-} \longrightarrow (D_N)^{\mathbb{Z}_-}$ is continuous with respect to the product topologies in $(D_d)^{\mathbb{Z}_-}$ and $(D_N)^{\mathbb{Z}_-}$.

As we show in the next few paragraphs, a given filter admits non-unique representations which can be generated with maps between state spaces resulting in systems in general with different memory and forecasting capacities. In the next proposition we show the important implication of the echo state property for our analysis. More specifically, we show that whenever the map between state spaces is a system morphism, the target system of the system morphism has the ESP and for the original system the existence of at least one solution for each input is guaranteed, then this solution is also unique or, equivalently, this system also has ESP and, moreover, the filters associated to these two systems are identical. This result can be made even stronger for isomorphisms which we will be using in our derivations.

State-space morphisms. The state-space representations of a given filter, when they exist, are not necessarily unique. These different realizations can be generated using maps between state spaces that satisfy certain natural functorial properties that make them into morphisms in the category of state-space systems. Additionally, as we see later on in Proposition 2.3, morphisms encode information about the solution properties and the echo state property of the systems that are linked by them. Consider the state-space systems determined by the two pairs (F_i, h_i), $i \in \{1, 2\}$, with $F_i : D_{N_i} \times D_d \longrightarrow D_{N_i}$ and $h_i : D_{N_i} \longrightarrow \mathbb{R}^m$.

Definition 2.2. A map $f : D_{N_1} \longrightarrow D_{N_2}$ is a ***morphism*** between the systems (F_1, h_1) and (F_2, h_2) whenever it satisfies the following two properties:

(i) ***System equivariance:*** $f(F_1(\mathbf{x}_1, \mathbf{z})) = F_2(f(\mathbf{x}_1), \mathbf{z})$, for all $\mathbf{x}_1 \in D_{N_1}$, and $\mathbf{z} \in D_d$.

(ii) ***Readout invariance:*** $h_1(\mathbf{x}_1) = h_2(f(\mathbf{x}_1))$, for all $\mathbf{x}_1 \in D_{N_1}$.

When the map f has an inverse f^{-1} and this inverse is also a morphism between the systems determined by the pairs (F_1, h_1) and (F_2, h_2) we say that f is a ***system isomorphism*** and that the systems (F_1, h_1) and (F_2, h_2) are ***isomorphic***. Given a system $F_1 : D_{N_1} \times D_d \longrightarrow D_{N_1}, h_1 : D_{N_1} \longrightarrow \mathbb{R}^m$ and a bijection $f : D_{N_1} \longrightarrow D_{N_2}$, the map f is a system isomorphism with respect to the system $F_2 : D_{N_2} \times D_d \longrightarrow D_{N_2}, h_2 : D_{N_2} \longrightarrow \mathbb{R}^m$ defined by

$$F_2(\mathbf{x}_2, \mathbf{z}) := f(F_1(f^{-1}(\mathbf{x}_2), \mathbf{z})), \quad \text{for all} \quad \mathbf{x}_2 \in D_{N_2}, \mathbf{z} \in D_d, \tag{2.2}$$

$$h_2(\mathbf{x}_2) := h_1(f^{-1}(\mathbf{x}_2)), \quad \text{for all} \quad \mathbf{x}_2 \in D_{N_2}. \tag{2.3}$$

Proposition 2.3. *Let (F_i, h_i), $i \in \{1, 2\}$, be two systems with $F_i : D_{N_i} \times D_d \longrightarrow D_{N_i}$ and $h_i : D_{N_i} \longrightarrow \mathbb{R}^m$. Let $f : D_{N_1} \longrightarrow D_{N_2}$ be a map. Then:*

(i) *If f is system equivariant and $\mathbf{x}^1 \in (D_{N_1})^{\mathbb{Z}_-}$ is a solution for the state system associated to F_1 and the input $\mathbf{z} \in (D_d)^{\mathbb{Z}_-}$, then so is $(f(\mathbf{x}^1_t))_{t\in\mathbb{Z}_-} \in (D_{N_2})^{\mathbb{Z}_-}$ for the system associated to F_2 and the same input.*

(ii) *Suppose that the system determined by (F_2, h_2) has the echo state property and assume that the state system determined by F_1 has at least one solution for each element $\mathbf{z} \in (D_d)^{\mathbb{Z}_-}$. If f is a morphism between (F_1, h_1) and (F_2, h_2), then (F_1, h_1) has the echo state property and, moreover,*

$$U^{F_1}_{h_1} = U^{F_2}_{h_2}. \tag{2.4}$$

(iii) *If f is a system isomorphism then the implications in the previous two points are reversible, that is, the indices 1 and 2 can be exchanged.*

Input and output stochastic processes. We now fix a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ on which all random variables are defined. The triple consists of the sample space Ω, which is the set of possible outcomes, the σ-algebra $\mathcal{A}$ (a set of subsets of Ω (events)), and a probability measure $\mathbb{P} : \mathcal{A} \longrightarrow [0, 1]$. The input signal is modeled as a discrete-time stochastic process $\mathbf{Z} = (\mathbf{Z}_t)_{t\in\mathbb{Z}_-}$ taking values in $D_d \subset \mathbb{R}^d$. Moreover, we write $\mathbf{Z}(\omega) = (\mathbf{Z}_t(\omega))_{t\in\mathbb{Z}_-}$ for each outcome $\omega \in \Omega$ to denote the realizations or sample paths of $\mathbf{Z}$. Since $\mathbf{Z}$ can be seen as a random sequence in $D_d \subset \mathbb{R}^d$, we write interchangeably $\mathbf{Z} : \mathbb{Z}_- \times \Omega \longrightarrow D_d$ and $\mathbf{Z} : \Omega \longrightarrow (D_d)^{\mathbb{Z}_-}$. The latter is by assumption measurable with respect to the Borel σ-algebra induced by the product topology in $(D_d)^{\mathbb{Z}_-}$. In this paper we consider only memory reconstruction and forecasting information processing tasks and hence the target process which is commonly denoted in the literature by $\mathbf{Y}$, will always be a time backward or forward shifted version of the input time series process $\mathbf{Z}$.

We will most of the time work under ***stationarity*** hypotheses. We recall that the discrete-time process $\mathbf{Z} : \Omega \longrightarrow (D_d)^{\mathbb{Z}_-}$ is stationary whenever $T_{-\tau}(\mathbf{Z}) \stackrel{\mathrm{d}}{=} \mathbf{Z}$, for any $\tau \in \mathbb{Z}_-$, where the symbol $\stackrel{\mathrm{d}}{=}$ stands for the equality in distribution and $T_{-\tau} : (\mathbb{R}^d)^{\mathbb{Z}_-} \longrightarrow (\mathbb{R}^d)^{\mathbb{Z}_-}$ is the ***time delay*** operator defined by $T_{-\tau}(\mathbf{z})_t := \mathbf{z}_{t+\tau}$ for any $t \in \mathbb{Z}_-$.

The definition of stationarity that we just formulated is usually known in the time series literature as ***strict stationarity*** [48]. When a process $\mathbf{Z}$ has second-order moments, that is, $\mathbf{Z}_t \in L^2(\Omega, \mathbb{R}^d)$, $t \in \mathbb{Z}_-$, then strict stationarity implies the so-called ***second-order stationarity***. We recall that a square-integrable process $\mathbf{Z} : \Omega \longrightarrow (D_d)^{\mathbb{Z}_-}$ is second-order stationary whenever **(i)** there exists a constant $\boldsymbol{\mu}_Z \in \mathbb{R}^d$ such that $\mathrm{E}[\mathbf{Z}_t] = \boldsymbol{\mu}_Z$, for all $t \in \mathbb{Z}_-$ (***mean stationarity***) and **(ii)** the autocovariance matrices $\mathrm{Cov}(\mathbf{Z}_t, \mathbf{Z}_{t+h})$ depend only on $h \in \mathbb{Z}$ and not on $t \in \mathbb{Z}_-$ (***autocovariance stationary***) and we can hence define the ***autocovariance function*** $\gamma : \mathbb{Z} \longrightarrow \mathbb{S}_d$ (with $\mathbb{S}_d$ the cone of positive semi-definite symmetric matrices of dimension d) as $\gamma(h) := \mathrm{Cov}(\mathbf{Z}_t, \mathbf{Z}_{t+h})$, with $t \in \mathbb{Z}_-$ arbitrarily chosen so that $t + h \in \mathbb{Z}_-$. The autocovariance function necessarily satisfies $\gamma(h) = \gamma(-h)^\top$ [48]. If $\mathbf{Z}$ is mean stationary and condition **(ii)** only holds for $h = 0$ we say that $\mathbf{Z}$ is ***covariance stationary***. Second-order stationarity and stationarity are only equivalent for Gaussian processes. If $\mathbf{Z}$ is autocovariance stationary and $\gamma(h) = 0$ for any non-zero $h \in \mathbb{Z}$ then we say that $\mathbf{Z}$ is a ***white noise***.

Corollary 2.4. *Let $F : D_N \times D_d \longrightarrow D_N$ be a state map that satisfies the hypotheses of Proposition 2.1 or that, more generally, has the echo state property and the associated filter $U^F : (D_d)^{\mathbb{Z}_-} \longrightarrow (D_N)^{\mathbb{Z}_-}$ is continuous with respect to the product topologies in $(D_d)^{\mathbb{Z}_-}$ and $(D_N)^{\mathbb{Z}_-}$. If the input process $\mathbf{Z} : \Omega \longrightarrow (D_d)^{\mathbb{Z}_-}$ is stationary, then so is the state $\mathbf{X} := U^F(\mathbf{Z}) : \Omega \longrightarrow (D_N)^{\mathbb{Z}_-}$ as well as the joint processes $(T_{-\tau}(\mathbf{X}), \mathbf{Z})$ and $(\mathbf{X}, T_{-\tau}(\mathbf{Z}))$, for any $\tau \in \mathbb{Z}_-$.*

In Proposition 2.3 we showed how to design alternative state-representations of a given filter by using state morphisms. This freedom can be put at work by choosing representations that have specific technical advantages that are needed in a given situation. An important implementation example of this strategy is the next Proposition, where we show that if we have a state system and an input for which the output process is covariance stationary and the corresponding covariance matrix is non-singular, then an isomorphic system representation exists whose corresponding state process is standardized, that is, the states have mean zero and covariance matrix equal to the identity. This standardization leads to systems that are easier to handle in terms of the computation of memory and forecasting capacities, which is profusely exploited later on in the main results of the paper.

Proposition 2.5 (*Standardization of State-space Realizations*). *Consider a state-space system as in (2.1a)–(2.1b) and suppose that the input process* $\mathbf{Z}:\Omega\longrightarrow(D_d)^{\mathbb{Z}_-}$ *is such that the associated state process* $\mathbf{X}:\Omega\longrightarrow(D_N)^{\mathbb{Z}_-}$ *is covariance stationary. Let* $\boldsymbol{\mu}:=\mathrm{E}[\mathbf{X}_t]$ *and suppose that the covariance matrix* $\Gamma_{\mathbf{X}}:=\mathrm{Cov}(\mathbf{X}_t,\mathbf{X}_t)$ *is non-singular. Then, the map* $f:\mathbb{R}^N\longrightarrow\mathbb{R}^N$ *given by* $f(\mathbf{x}):=\Gamma_{\mathbf{X}}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})$ *is a system isomorphism between the system (2.1a)–(2.1b) and the one with state map*

$$\widetilde{F}(\mathbf{x},\mathbf{z}):=\Gamma_{\mathbf{X}}^{-1/2}\left(F\left(\Gamma_{\mathbf{X}}^{1/2}\mathbf{x}+\boldsymbol{\mu},\mathbf{z}\right)-\boldsymbol{\mu}\right) \tag{2.5}$$

and readout

$$\widetilde{h}(\mathbf{x}):=(\Gamma_{\mathbf{X}}^{-1/2}\mathbf{W})^{\top}\mathbf{x}+\mathbf{W}^{\top}\boldsymbol{\mu}+\mathbf{a}. \tag{2.6}$$

Moreover, the state process $\widetilde{\mathbf{X}}$ *associated to the system* $\widetilde{F}$ *and the input* $\mathbf{Z}$ *is covariance stationary and*

$$\mathrm{E}[\widetilde{\mathbf{X}}_t]=\mathbf{0},\quad \textit{and}\quad \mathrm{Cov}(\widetilde{\mathbf{X}}_t,\widetilde{\mathbf{X}}_t)=\mathbb{I}_N. \tag{2.7}$$

3. Memory and forecasting capacity bounds for stationary inputs

The following definition extends the notion of memory capacity introduced in [1] to general nonlinear systems and to input signals that are stationary but not necessarily time-decorrelated.

Definition 3.1. Let $\mathbf{Z}:\Omega\longrightarrow D^{\mathbb{Z}_-}$, $D\subset\mathbb{R}$, be a variance-stationary input and let F be a state map that has the echo state property with respect to the paths of $\mathbf{Z}$. Assume, moreover, that the associated state process $\mathbf{X}:\Omega\longrightarrow(D_N)^{\mathbb{Z}_-}$ defined by $\mathbf{X}_t:=U^F(\mathbf{Z})_t$ is covariance stationary, as well as the joint processes $(T_{-\tau}(\mathbf{X}),\mathbf{Z})$ and $(\mathbf{X},T_{-\tau}(\mathbf{Z}))$, for any $\tau\in\mathbb{Z}_-$. We define the τ-lag ***memory capacity*** MC_τ (respectively, ***forecasting capacity*** FC_τ) of F with respect to $\mathbf{Z}$ as:

$$\mathrm{MC}_\tau:=1-\frac{1}{\mathrm{Var}(Z_t)}\min_{\substack{\mathbf{W}\in\mathbb{R}^N\\ a\in\mathbb{R}}}\mathrm{E}\left[\left((T_{-\tau}\mathbf{Z})_t-\mathbf{W}^{\top}U^F(\mathbf{Z})_t-a\right)^2\right], \tag{3.1}$$

$$\mathrm{FC}_\tau:=1-\frac{1}{\mathrm{Var}(Z_t)}\min_{\substack{\mathbf{W}\in\mathbb{R}^N\\ a\in\mathbb{R}}}\mathrm{E}\left[\left(Z_t-\mathbf{W}^{\top}U^F(T_{-\tau}(\mathbf{Z}))_t-a\right)^2\right]. \tag{3.2}$$

The ***total memory capacity*** MC (respectively, ***total forecasting capacity*** FC) of F with respect to $\mathbf{Z}$ is defined as:

$$\mathrm{MC}:=\sum_{\tau\in\mathbb{Z}_-}\mathrm{MC}_\tau,\qquad \mathrm{FC}:=\sum_{\tau\in\mathbb{Z}^-}\mathrm{FC}_\tau. \tag{3.3}$$

Note that, by Corollary 2.4, the conditions of this definition are met when, for instance, U^F is continuous with respect to the product topologies and the input process $\mathbf{Z}$ is stationary.

The optimization problems appearing in the definitions (3.1) and (3.2) of the memory and forecasting capacities can be explicitly solved when the state covariance matrix $\Gamma_{\mathbf{X}}:=\mathrm{Cov}(\mathbf{X}_t,\mathbf{X}_t)$ is invertible. We refer to this situation as the *regular case*. These solutions are provided in the following lemma.

Lemma 3.2. *In the conditions of Definition 3.1 and if the covariance matrix* $\Gamma_{\mathbf{X}}:=\mathrm{Cov}(\mathbf{X}_t,\mathbf{X}_t)$ *is invertible then, for any* $\tau\in\mathbb{Z}_-$*:*

$$\begin{aligned}\mathrm{MC}_\tau&=\frac{\mathrm{Cov}\left(Z_{t+\tau},\mathbf{X}_t\right)\Gamma_{\mathbf{X}}^{-1}\mathrm{Cov}\left(\mathbf{X}_t,Z_{t+\tau}\right)}{Var\left(Z_t\right)},\\ \mathrm{FC}_\tau&=\frac{\mathrm{Cov}\left(Z_t,\mathbf{X}_{t+\tau}\right)\Gamma_{\mathbf{X}}^{-1}\mathrm{Cov}\left(\mathbf{X}_{t+\tau},Z_t\right)}{Var\left(Z_t\right)}.\end{aligned} \tag{3.4}$$

The availability of the closed-form solutions in the expressions (3.1) and (3.2) make the computation of capacities much easier. This will become particularly evident in the next section devoted to linear systems. We emphasize that even for those simpler system specifications, the non-invertibility of the associated covariance matrices of states leads to technical difficulties. The same holds even to a greater extent for nonlinear systems. Some of those problems can be handled by using equivalent state-space representations. The next result shows that new representations obtained out of linear injective system morphisms leave invariant the capacities and hence can be used to produce systems with more technically tractable properties. This result will be used later on in Section 4 when we study the memory and forecasting capacities of linear systems in the singular case.

Lemma 3.3. *Let* $\mathbf{Z}$ *be a variance-stationary input and let* $F_2:D_{N_2}\times D\longrightarrow D_{N_2}$ *be a state map that satisfies the conditions of Definition 3.1. Let* $F_1:D_{N_1}\times D\longrightarrow D_{N_1}$ *be another state map that has at least one solution for each* $\mathbf{z}\in D^{\mathbb{Z}_-}$ *and let* $f:\mathbb{R}^{N_1}\longrightarrow\mathbb{R}^{N_2}$ *be an injective linear system equivariant map between* F_1 *and* F_2*. Then, the memory and forecasting capacities of* F_1 *with respect to* $\mathbf{Z}$ *are well-defined and coincide with those of* F_2 *with respect to* $\mathbf{Z}$*.*

The next theorem is the first main contribution of this paper and generalizes the bounds formulated in [1] for the total memory capacity of an echo state network in the presence of independent inputs to general state systems with second-order stationary inputs and invertible state covariance matrices. We show that both the total memory and forecasting capacities are nonnegative and that upper bounds can be formulated that are fully determined by the behavior of the autocovariance or the spectral density functions of the input and the dimensionality of the state space.

Theorem 3.4. *Suppose that we are in the conditions of Definition 3.1 and that the covariance matrix* $\Gamma_{\mathbf{X}}:=\mathrm{Cov}(\mathbf{X}_t,\mathbf{X}_t)$ *is non-singular.*

(i) *For any* $\tau\in\mathbb{Z}_-$*:*

$$0\leq\mathrm{MC}_\tau\leq1\quad\textit{and}\quad 0\leq\mathrm{FC}_\tau\leq1. \tag{3.5}$$

(ii) *Suppose that, additionally, the input process* $\mathbf{Z}$ *is second-order stationary with autocovariance function* $\gamma:\mathbb{Z}\longrightarrow\mathbb{R}$*, and that for any* $L\in\mathbb{Z}_-$ *the symmetric matrices* $H^L\in\mathbb{M}_{-L+1}$ *determined by* $H^L_{ij}:=\gamma(|i-j|)$ *are invertible. Then, if we use the symbol* C *to denote both* MC *and* FC *in (3.3), we have:*

$$0\leq \mathrm{C}\leq\frac{N}{\gamma(0)}\rho(H)\leq N\left(1+\frac{2}{\gamma(0)}\sum_{j=1}^{\infty}|\gamma(j)|\right), \tag{3.6}$$

where $\rho(H):=\lim_{L\to-\infty}\rho(H^L)$*, with* $\rho(H^L)$ *the spectral radius of* H^L*.*

(iii) *In the same conditions as in part* **(ii)***, suppose that, additionally, the autocovariance function* γ *is absolutely summable, that* *is, $\sum_{j=-\infty}^{\infty}|\gamma(j)| < +\infty$. In that case, the spectral density $f:[-\pi,\pi]\to\mathbb{R}$ of $\mathbf{Z}$ is well-defined and given by*

$$f(\lambda) := \frac{1}{2\pi}\sum_{n\in\mathbb{Z}} e^{-in\lambda}\gamma(n), \tag{3.7}$$

and we have that

$$0 \leq \mathrm{C} \leq \frac{2\pi N}{\gamma(0)}M_f \leq N\left(1+\frac{2}{\gamma(0)}\sum_{j=1}^{\infty}|\gamma(j)|\right), \tag{3.8}$$

where $M_f := \max_{\lambda\in[-\pi,\pi]}\{f(\lambda)\}$.

We emphasize that the results in this theorem hold for a general class of nonlinear recurrent neural networks with linear readouts both with trainable or with randomly generated neuron weights (reservoir computing). Additionally, it can be used as a tool in the design of the network architecture when the autocovariance structure of the input is known. These bounds show in passing that memory and forecasting capacities are determined not only by a system but also to a great extent by the memory of the input process itself.

This general result and its proof are used in the next corollary to recover the total memory capacity bounds proposed in [1] when using independent inputs and to show that, in that case, the total forecasting capacity is always zero.

Corollary 3.5 ([1]). *In the conditions of Theorem 3.4, if the inputs $\{Z_t\}_{t\in\mathbb{Z}_-}$ are independent, then*

$$0 \leq \mathrm{MC} \leq N \quad \text{and} \quad \mathrm{FC} = 0. \tag{3.9}$$

The proofs of the previous two results, which can be found in the Appendix, shed some light on the relative values of the forecasting and memory capacities, as well as on the quality of the common bounds in (3.6) and (3.8). Indeed, these estimates are obtained by finding upper bounds for the norms of the orthogonal projections (in the L^2 sense) of the state at a given time on the vector space generated by all the inputs fed into the system up until that point in time, in the case of the memory capacity and, for the forecasting capacity, by the inputs that will be fed in the future. The built-in causality of state-space filters implies that the state has a functional dependence exclusively on past inputs and hence its projection onto future inputs becomes nontrivial only via dependence phenomena in the input signal. This fact brings in its wake that, typically, even in the presence of strongly autocorrelated input signals, the projection of the state vector onto past inputs produces larger vectors (in norm) than onto future ones. The bounding mechanism used in the proof (see (A.13)) is not able to take this fact into account, as that would entail using specific knowledge on the functional form of the filter, which is something that we avoided in the pursuit of generic bounds that are common to all state-space systems with a given dimension. The price to pay for this degree of generality is that the bounds will be closer to the memory than to the forecasting capacities and hence will be sharper for the former than for the latter. Later on in Section 4.3 we illustrate these facts with a numerical example and we additionally discuss the sharpness question, or rather the lack of it, in the presence of dependent inputs.

4. The memory and forecasting capacities of linear systems

When the state equation (2.1a) is linear and has the echo state property, both the memory and forecasting capacities in (3.3) can be explicitly written down in terms of the equation parameters provided that the invertibility hypothesis on the covariance matrix of the states holds. This case has been studied for independent inputs and randomly generated linear systems in [1,10] and, more recently, in [27] for more general correlated inputs and diagonalizable linear systems. It is in this paper that it has been pointed out for the first time how different linear systems that maximize forecasting and memory capacities may be.

This section contains the second main contribution of the paper. We split it in two parts. In the first one we handle what we call the regular case in which we assume the invertibility of the covariance matrix of the states. In the second one we see how, using system morphisms of the type introduced in Lemma 3.3, we can reduce the general singular case to the regular one. This approach allows us to prove that *when the inputs are independent, the memory capacity of a linear system with independent inputs coincides with the rank of its associated controllability or reachability matrix*. A rigorous proof of this result is, to our knowledge, not available in the literature. This statement is a generalization of the fact, already established in [1], that when the rank of the controllability matrix is maximal, then the linear system has maximal capacity, that is, its capacity coincides with the dimensionality of its state space. Different configurations that maximize the rank of the controllability matrix have been recently studied in [32,33]. We find the results in this section useful also from the applications point of view as they allow the design of linear recurrent networks with an exact pre-specified memory capacity.

4.1. The regular case

The explicit capacity formulas that we state in the next result only require the stationarity of the input and the invertibility of the covariance matrix of the states. In this case we consider fully infinite inputs, that is, we work with a stationary input process $\mathbf{Z}:\Omega\longrightarrow\mathbb{R}^{\mathbb{Z}}$ that has second-order moments and an autocovariance function $\gamma:\mathbb{Z}\longrightarrow\mathbb{R}$ for which we construct the semi-infinite (respectively, doubly infinite) symmetric positive semi-definite Toeplitz matrix $H_{ij} := \gamma(|i-j|)$, $i,j\in\mathbb{N}^+$ (respectively, $\overline{H}_{ij} := \gamma(|i-j|)$, $i,j\in\mathbb{Z}$).

Proposition 4.1. *Consider the linear state system determined by the linear state map $F:\mathbb{R}^N\times\mathbb{R}\longrightarrow\mathbb{R}^N$ given by*

$$F(\mathbf{x},z) := A\mathbf{x}+\mathbf{C}z, \quad \text{with } \mathbf{C}\in\mathbb{R}^N, A\in\mathbb{M}_N, |||A||| = \sigma_{\max}(A) < 1, \tag{4.1}$$

where $\sigma_{\max}(A)$ is the largest singular value of the matrix A (usually referred to as ***connectivity matrix****). Let $D\subset\mathbb{R}$ be compact and consider a zero-mean stationary input process $\mathbf{Z}:\Omega\longrightarrow D^{\mathbb{Z}}$ that has second-order moments and an absolutely summable autocovariance function $\gamma:\mathbb{Z}\longrightarrow\mathbb{R}$. Suppose also that the associated spectral density f satisfies that $f(\lambda)\geq 0$, $\lambda\in[-\pi,\pi]$, and that $f(\lambda)=0$ holds only in at most a countable number of points.*

Then F has the echo state property and the associated filter $U^{A,\mathbf{C}}$ is such that its output $\mathbf{X} := U^{A,\mathbf{C}}(\mathbf{Z}):\Omega\longrightarrow(D_N)^{\mathbb{Z}_-}$ as well as the joint processes $(T_{-\tau}(\mathbf{X}),\mathbf{Z})$ and $(\mathbf{X},T_{-\tau}(\mathbf{Z}))$, for any $\tau\in\mathbb{Z}_-$, are stationary and $\mathbf{X}$ is covariance stationary. Suppose that the covariance matrix $\Gamma_{\mathbf{X}} := \mathrm{Cov}(\mathbf{X}_t,\mathbf{X}_t)$ is non-singular. Then,

(i) *Consider the N vectors $\mathbf{B}_1,\ldots,\mathbf{B}_N\in\ell^2_+(\mathbb{R})$ defined by*

$$B_i^j := \left(\Gamma_{\mathbf{X}}^{-1/2}\sum_{k=0}^{\infty}A^k\mathbf{C}H_{k+1,j}^{1/2}\right)_i, \quad j\in\mathbb{N}^+,\ i=1,\ldots,N, \tag{4.2}$$

where the square root matrices are computed via orthogonal diagonalization. These entries are all finite and form vectors *that constitute an orthonormal set in* $\ell^2_+(\mathbb{R})$*. The total memory capacity* MC *can be written as*

$$\mathrm{MC} = \frac{1}{\gamma(0)} \sum_{i=1}^{N} \langle \mathbf{B}_i, H\mathbf{B}_i \rangle_{\ell^2}. \tag{4.3}$$

(ii) *Consider the N vectors* $\mathbf{B}_1, \ldots, \mathbf{B}_N \in \ell^2(\mathbb{R})$ *defined by*

$$B_i^j := \left(\Gamma_{\mathbf{X}}^{-1/2} \sum_{k=0}^{\infty} A^k \mathbf{C} \overline{H}_{-k,j}^{1/2} \right)_i, \quad j \in \mathbb{Z},\ i = 1, \ldots, N. \tag{4.4}$$

These vectors form an orthonormal set in $\ell^2(\mathbb{R})$ *and the forecasting capacity* FC *can be written as*

$$\mathrm{FC} = \frac{1}{\gamma(0)} \sum_{i=1}^{N} \left\| \mathbb{P}_{\mathbb{Z}^+} \left(\overline{H}^{1/2} \mathbf{B}_i \right) \right\|_{\ell^2}^2, \tag{4.5}$$

where $\mathbb{P}_{\mathbb{Z}^+} : \ell^2(\mathbb{R}) \longrightarrow \ell^2(\mathbb{R})$ *is the projection that sets to zero all the entries with non-positive index.*

When the inputs are second-order stationary and not autocorrelated ($\mathbf{Z}$ is a white noise) then the formulas in the previous result can be used to give, for the linear case, a more informative version of Corollary 3.5, without the need to invoke input independence.

Corollary 4.2. *Suppose that we are in the hypotheses of Proposition 4.1 and that, additionally, the input process* $\mathbf{Z} : \Omega \longrightarrow D^{\mathbb{Z}}$ *is a white noise, that is, the autocovariance function* γ *satisfies that* $\gamma(h) = 0$*, for any non-zero* $h \in \mathbb{Z}$*. Then:*

$$\mathrm{MC} = N \quad \text{and} \quad \mathrm{FC} = 0. \tag{4.6}$$

An important conclusion of this corollary is that linear systems with white noise inputs that have a non-singular state covariance matrix $\Gamma_{\mathbf{X}}$ automatically have maximal memory capacity. This makes important the characterization of the invertibility of $\Gamma_{\mathbf{X}}$ in terms of the parameters $A \in \mathbb{M}_N$ and $\mathbf{C} \in \mathbb{R}^N$ in (4.1). The following proposition provides such a characterization, which has serious practical implications at the time of designing linear recurrent networks, and establishes a connection between the invertibility of $\Gamma_{\mathbf{X}}$ and Kalman's characterization of the controllability of a linear system [28].

Proposition 4.3. *Consider the linear state system introduced in (4.1). Suppose that the input process* $\mathbf{Z} : \Omega \longrightarrow D^{\mathbb{Z}}$ *is a white noise and that the connectivity matrix* $A \in \mathbb{M}_N$ *is diagonalizable. Let* $\sigma(A) = \{\lambda_1, \ldots, \lambda_N\}$ *be the spectrum of A and let* $\{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$ *be an eigenvectors basis.*

(i) *The state covariance matrix* $\Gamma_{\mathbf{X}} = \gamma(0) \sum_{j=0}^{\infty} A^j \mathbf{C}\mathbf{C}^\top \left(A^j\right)^\top$ *is non-singular if and only if all the eigenvalues in* $\sigma(A)$ *are distinct.*

(ii) *The vectors* $\left\{A\mathbf{C}, A^2\mathbf{C}, \ldots, A^N\mathbf{C}\right\}$ *form a basis of* $\mathbb{R}^N$ *if and only if all the eigenvalues in* $\sigma(A)$ *are distinct and non-zero and in the linear decomposition* $\mathbf{C} = \sum_{i=1}^{N} c_i \mathbf{v}_i$*, the coefficients* c_i*,* $i \in \{1, \ldots, N\}$ *are all non-zero.*

(iii) *The conditions in the previous point are equivalent to the* **Kalman controllability condition***: the vectors* $\{\mathbf{C}, A\mathbf{C}, \ldots, A^{N-1}\mathbf{C}\}$ *form a basis of* $\mathbb{R}^N$ *together with the condition that all the eigenvalues in* $\sigma(A)$ *are non-zero.*

It has been shown in [1] that the controllability condition is equivalent to maximum memory capacity. The approach followed in this proposition will allow us in Theorem 4.4 to generalize this statement by proving that the memory capacity *equals* the rank of the controllability matrix (introduced in detail later on in the text).

4.2. The singular case

In the following paragraphs we study the situation in which the covariance matrix of the states $\Gamma_{\mathbf{X}}$ with white noise inputs is not invertible. All the results formulated in the paper so far, in particular the capacity formulas in (3.4), are not valid anymore in this case. However, it is well-known that the non-invertibility of the covariance matrix of the states process for systems with high state space dimensionality is a frequent issue. In the reservoir computing literature this problem is usually overcome at the time of training via the use of spectral regularization techniques, like for instance the Tikhonov regularized regressions. In this paper we adopt a different strategy to tackle this problem, namely we use the idea introduced in Lemma 3.3 of using system morphisms that leave capacities invariant. More specifically, we show that whenever we are given a linear system whose covariance matrix $\Gamma_{\mathbf{X}}$ is not invertible, there exists another linear system defined in a dimensionally smaller state space that generates the same filter and hence has the same capacities but, unlike the original system, this smaller one has an invertible covariance matrix. This feature allows us to use for this system some of the results in previous sections and, in particular *to compute its memory capacity in the presence of independent inputs* that, as we establish in the next theorem, *coincides with the rank of the controllability matrix.*

Theorem 4.4. *Consider the linear system* $F(\mathbf{x}, z) := A\mathbf{x} + \mathbf{C}z$ *introduced in (4.1) and suppose that the input process* $\mathbf{Z} : \Omega \longrightarrow D^{\mathbb{Z}}$ *is a strictly stationary white noise. Let* $R(A, \mathbf{C}) := \left(\mathbf{C}|A\mathbf{C}|\cdots|A^{N-1}\mathbf{C}\right)$ *be the controllability matrix of the linear system.*

(i) *Then it holds that*

$$\ker \Gamma_{\mathbf{X}} = \ker R(A, \mathbf{C})^\top. \tag{4.7}$$

(ii) *Let* $X := \mathrm{span}\left\{\mathbf{C}, A\mathbf{C}, \ldots, A^{N-1}\mathbf{C}\right\}$ *and let* $r := \dim(X) = \mathrm{rank}\, R(A, \mathbf{C})$*. Let* $V \subset \mathbb{R}^N$ *be a vector subspace such that* $\mathbb{R}^N = X \oplus V$ *and let* $i_X : X \hookrightarrow \mathbb{R}^N$ *and* $\pi_X : \mathbb{R}^N \longrightarrow X$ *be the injection and the projection associated to this splitting, respectively. Then, the linear system* $\overline{F} : X \times D \longrightarrow X$ *defined by*

$$\overline{F}(\overline{\mathbf{x}}, z) := \overline{A}\overline{\mathbf{x}} + \overline{\mathbf{C}}z \text{ and determined by } \overline{A} := \pi_X A i_X \text{ and } \overline{\mathbf{C}} = \pi_X(\mathbf{C}), \tag{4.8}$$

is well-defined and has the echo state property.

(iii) *In the notation of part* **(ii)***, the map* $i_X : X \hookrightarrow \mathbb{R}^N$ *is an injective linear system equivariant map between* $\overline{F}$ *and F.*

(iv) *In the notation of part* **(ii)***, let* $\overline{\mathbf{X}} : \Omega \longrightarrow X^{\mathbb{Z}}$ *be the output of the filter determined by the state-system* $\overline{F}$*. Then,*

$$\mathrm{rank}\, R(\overline{A}, \overline{\mathbf{C}}) = \mathrm{rank}\, R(A, \mathbf{C}), \tag{4.9}$$

and if $\overline{A}$ *is diagonalizable with non-zero eigenvalues then* $\Gamma_{\overline{\mathbf{X}}} := \mathrm{Cov}(\overline{\mathbf{X}}_t, \overline{\mathbf{X}}_t)$ *is invertible.*

(v) *If* $\overline{A}$ *is diagonalizable with non-zero eigenvalues then the memory* MC *and forecasting* FC *capacities of F with respect to* $\mathbf{Z}$ *are given by*

$$\mathrm{MC} = \mathrm{rank}\, R(A, \mathbf{C}) = \dim\left(\mathrm{span}\left\{\mathbf{C}, A\mathbf{C}, \ldots, A^{N-1}\mathbf{C}\right\}\right) \quad \text{and} \quad \mathrm{FC} = 0. \tag{4.10}$$

The statement in part **(v)** provides a generalization of the statement in [1] that establishes the equivalence between controllability ($\mathrm{rank}\, R(A, \mathbf{C}) = N$) and maximum memory capacity ($\mathrm{MC} = N$). More specifically, our statement shows that the memory capacity *equals* the rank of the controllability matrix of the linear system. This result has far reaching implications for the applications of recurrent linear networks with either fully trainable or just randomly generated weights. Given a precise computational task at hand, a learner can use the rank of the controllability matrix in order to construct a network with a prescribed memory capacity. In particular, in the reservoir computing community the result in [1] has deserved much attention. Since the controllability condition was known to be equivalent to maximal memory capacity, many attempts have been made trying to propose a strategy to generate random reservoirs which would have maximal expected controllability matrix rank [31–33,49]. Our results show that the same work can be done in those cases when one is interested in constructing random reservoirs with a required controllability matrix rank and that, as we proved, amounts to the memory capacity of the system.

4.3. Numerical illustration

An important consequence of part **(v)** in Theorem 4.4 is that the capacity bounds in (3.6) and (3.8) are sharp in the presence of independent inputs or, equivalently, that the bounds in Corollary 3.5 are sharp. Indeed, by (4.10), any linear system with independent inputs whose controllability matrix has maximal rank has full memory capacity equal to its dimension, and hence it achieves the upper bound in Corollary 3.5. A natural question that arises is if this sharpness remains valid for correlated inputs. Even though here we do not formulate general conditions that would ensure that fact, the following paragraphs contain a numerical illustration that give indications of what the situation may be. Indeed, we demonstrate that, in general, *the controllability condition does not ensure anymore the sharpness of the bounds* (3.6) and (3.8) *with correlated inputs* neither for memory capacities nor for forecasting capacities. The latter is a consequence of the arguments in the paragraph after Corollary 3.5.

The panels in Fig. 1 show (in logarithmic scale) numerically computed memory and forecasting capacities, as well as the bounds in (3.6) based on the spectral radius $\rho(H)$, for a linear system as in (4.1). In this experiment we chose $N = 15$ and a connectivity matrix A (spectral radius equal to 0.9) and an input mask vector $\mathbf{C}$ such that $\operatorname{rank} R(A, \mathbf{C}) = N = 15$. The resulting system has hence memory capacity equal to 15 in the presence of independent inputs.

This system has been then presented with three different types of autocorrelated inputs that are realizations of AR(1), MA(1), and ARMA(1,1) processes (see, for instance, [48] for details on these models) driven by independent standard normal innovations. We denote (as in the figure) by ϕ and θ the autoregressive and the moving-average coefficients needed in the specification of these models. The top two panels in Fig. 1 have been obtained by varying the values of ϕ (for the AR(1) case) and of θ (for the MA(1) case) between 0 and 1. In the one at the bottom we took ϕ equal to θ and we then varied them simultaneously between 0 and 1.

The curves in the figures show how the bounds and the capacities evolve as a function of those parameters. The capacities have been computed using the definition in (3.3) and the formulas (3.4) where we truncated the infinite sum at the value $\tau = 250$ and the covariances were empirically estimated using realizations of length ten thousand. In both cases, the values $\phi = 0$ and $\theta = 0$ correspond to the independent inputs case and the figures show how then the theoretical bounds correspond to the actual memory capacity of the system equal to 15. As soon as both parameters are non-zero we see in the figures a monotonous increase in the memory and forecasting capacities of the system, which is specially visible when it comes to the relation between the autoregressive parameter ϕ and the forecasting capacity. The figures also show that, as we anticipated, the theoretical bounds are strictly above the memory capacity even though we are in the presence of a system with full rank controllability matrix. This shows, in passing, that the results in Theorem 4.4 do not automatically extend to the dependent inputs case.

5. Conclusions

In this paper we have studied memory and forecasting capacities of generic nonlinear recurrent networks with respect to arbitrary stationary inputs that are not necessarily independent. In particular, we have stated *upper bounds for total memory and forecasting capacities in terms of the dimensionality of the network and the autocovariance of the inputs* that generalize those formulated in [1] for independent inputs.

The approach followed in the paper is particularly advantageous for linear networks for which explicit expressions can be formulated for both capacities. In the *classical* linear case with independent inputs, we have proved that *the memory capacity of a linear recurrent network with independent inputs is given by the rank of its controllability matrix*. This explicit and readily computable characterization of the memory capacity of those networks generalizes a well-known relation between maximal capacity and Kalman's controllability condition, formulated for the first time in [1]. This is, to our knowledge, the first rigorous proof of the relation between network memory and the rank of the controllability matrix, that has been for a long time part of the reservoir computing folklore.

The results in this paper suggest links between controllability and memory capacity for nonlinear recurrent systems that will be explored in forthcoming works.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank the editor and two anonymous referees for their careful reading and interesting comments that have improved the paper. L Gonon and JPO acknowledge partial financial support coming from the Research Commission of the Universität Sankt Gallen and the Swiss National Science Foundation (grant number 200021_175801/1). JPO acknowledges partial financial support of the French ANR "BIPHOPROC" project (ANR-14-OHRI-0002-02). The authors thank the hospitality and the generosity of the FIM at ETH Zurich and the Division of Mathematical Sciences of the Nanyang Technological University, Singapore, where a significant portion of the results in this paper were obtained.

Appendix

A.1. Proof of Proposition 2.3

(i) By hypothesis $\mathbf{x}_t^1 = F_1(\mathbf{x}_{t-1}^1, \mathbf{z}_t)$, for all $t \in \mathbb{Z}_-$. By system equivariance,

$$f(\mathbf{x}_t^1) = f(F_1(\mathbf{x}_{t-1}^1, \mathbf{z}_t)) = F_2(f(\mathbf{x}_{t-1}^1), \mathbf{z}_t), \quad \text{for all } t \in \mathbb{Z}_-,$$

as required.

(ii) In order to show that (2.4) holds, it suffices to prove that given any $\mathbf{z} \in (D_d)^{\mathbb{Z}_-}$, any solution $(\mathbf{y}^1, \mathbf{x}^1) \in (\mathbb{R}^m)^{\mathbb{Z}_-} \times \left(D_{N_1}\right)^{\mathbb{Z}_-}$ of (F_1, h_1) associated to $\mathbf{z}$, with $\mathbf{y}^1 := \left(h_1(\mathbf{x}_t^1)\right)_{t \in \mathbb{Z}_-}$, which exists by

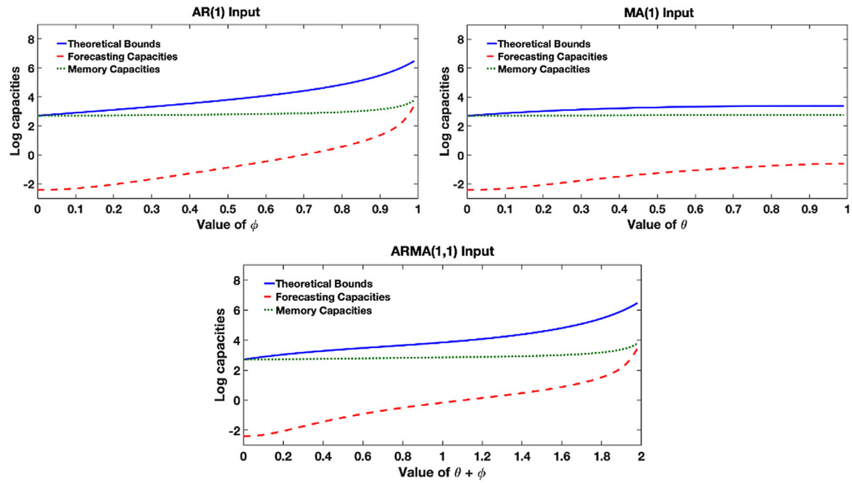

Fig. 1. Numerically computed memory and forecasting capacities of a linear system with full rank controllability matrix and AR(1), MA(1), and ARMA(1,1) inputs. The curves depict the behavior of the numerically computed capacities and the bounds in (3.6) when the input model parameters are varied. These results show that, in this case, the theoretical bounds are only sharp for independent inputs.

the hypothesis on F_1, coincides with the unique solution $U^{F_2}_{h_2}(\mathbf{z})$ for the system (F_2, h_2). Indeed, for any $t \in \mathbb{Z}_-$,

$$\mathbf{y}^1_t = h_1(F_1(\mathbf{x}^1_{t-1}, \mathbf{z}_t)) = h_2(f(F_1(\mathbf{x}^1_{t-1}, \mathbf{z}_t))) = h_2(F_2(f(\mathbf{x}^1_{t-1}), \mathbf{z}_t)).$$

Here, the second equality follows from readout invariance and the third one from the system equivariance. This implies that $\left(\mathbf{y}^1, (f(\mathbf{x}^1_t))_{t\in\mathbb{Z}_-}\right)$ is a solution of the system determined by (F_2, h_2) for the input $\mathbf{z}$. By hypothesis, (F_2, h_2) has the echo state property and hence $\mathbf{y}^1 = U^{F_2}_{h_2}(\mathbf{z})$ and since $\mathbf{z} \in (D_d)^{\mathbb{Z}_-}$ is arbitrary, the result follows. Part **(iii)** is straightforward. ■

A.2. Proof of Corollary 2.4

The continuity (and hence the measurability) hypothesis on U^F proves that $\mathbf{X} := U^F(\mathbf{Z})$ is stationary (see [50, page 157]). The joint processes $(T_{-\tau}(\mathbf{X}), \mathbf{Z})$ and $(\mathbf{X}, T_{-\tau}(\mathbf{Z}))$ are also stationary as they are the images of $\mathbf{Z}$ by the measurable maps $(T_{-\tau} \circ U^F) \times \mathbb{I}_{(\mathbb{R}^d)^{\mathbb{Z}_-}}$ and $U^F \times T_{-\tau}$, respectively. ■

A.3. Proof of Proposition 2.5

Since by hypothesis the matrix $\Gamma_{\mathbf{X}}$ is invertible, then so is its square root $\Gamma_{\mathbf{X}}^{1/2}$, as well as the map f, whose inverse f^{-1} is given by $f^{-1}(\mathbf{x}) := \Gamma_{\mathbf{X}}^{1/2}\mathbf{x} + \boldsymbol{\mu}$. The fact that f is a system isomorphism between (F, h) and $(\widetilde{F}, \widetilde{h})$ is a consequence of the equalities (2.2)–(2.3). Parts **(i)** and **(iii)** of Proposition 2.3 guarantee that if $\mathbf{X}$ is the state process associated to F and input $\mathbf{Z}$ then so is $\widetilde{\mathbf{X}}$ defined by $\widetilde{\mathbf{X}}_t := \Gamma_{\mathbf{X}}^{-1/2}(\mathbf{X}_t - \boldsymbol{\mu})$, $t \in \mathbb{Z}_-$, with respect to $\widetilde{F}$. The equalities (2.7) immediately follow. ■

A.4. Proof of Lemma 3.2

First of all, for any $\tau \in \mathbb{Z}_-$, consider the optimization problems

$$\begin{aligned}(\widehat{\mathbf{W}}_{\mathrm{MC}_\tau}, \hat{a}_{\mathrm{MC}_\tau}) &= \underset{\substack{\mathbf{W}\in\mathbb{R}^N\\ a\in\mathbb{R}}}{\arg\min}\, \mathrm{E}\left[\left((T_{-\tau}\mathbf{Z})_t - \mathbf{W}^\top U^F(\mathbf{Z})_t - a\right)^2\right]\\ &= \underset{\substack{\mathbf{W}\in\mathbb{R}^N\\ a\in\mathbb{R}}}{\arg\min}\, \mathrm{E}\left[\left(Z_{t+\tau} - \mathbf{W}^\top \mathbf{X}_t - a\right)^2\right] \qquad \text{(A.1)}\end{aligned}$$

and

$$\begin{aligned}(\widehat{\mathbf{W}}_{\mathrm{FC}_\tau}, \hat{a}_{\mathrm{FC}_\tau}) &= \underset{\substack{\mathbf{W}\in\mathbb{R}^N\\ a\in\mathbb{R}}}{\arg\min}\, \mathrm{E}\left[\left(Z_t - \mathbf{W}^\top U^F(T_{-\tau}(\mathbf{Z}))_t - a\right)^2\right]\\ &= \underset{\substack{\mathbf{W}\in\mathbb{R}^N\\ a\in\mathbb{R}}}{\arg\min}\, \mathrm{E}\left[\left(Z_t - \mathbf{W}^\top \mathbf{X}_{t+\tau} - a\right)^2\right] \qquad \text{(A.2)}\end{aligned}$$

in (3.1) and (3.2) of Definition 3.1, respectively. It is straightforward to prove by setting equal to zero the derivatives of the objective functions with respect to the optimized parameters (see Section C in the Technical Supplement of [15] for details) that both (A.1) and (A.2) admit closed-form solutions when $\Gamma_{\mathbf{X}}$ is invertible and they are given by

$$\begin{aligned}\widehat{\mathbf{W}}_{\mathrm{MC}_\tau} &= \mathrm{Cov}(\mathbf{X}_t, \mathbf{X}_t)^{-1}\mathrm{Cov}(\mathbf{X}_t, Z_{t+\tau}) = \Gamma_{\mathbf{X}}^{-1}\mathrm{Cov}(\mathbf{X}_t, Z_{t+\tau}),\\ \hat{a}_{\mathrm{MC}_\tau} &= \mu_Z - \widehat{\mathbf{W}}_{\mathrm{MC}_\tau}^\top \boldsymbol{\mu}\end{aligned}$$

and

$$\begin{aligned}\widehat{\mathbf{W}}_{\mathrm{FC}_\tau} &= \mathrm{Cov}(\mathbf{X}_{t+\tau}, \mathbf{X}_{t+\tau})^{-1}\mathrm{Cov}(\mathbf{X}_{t+\tau}, Z_t) = \Gamma_{\mathbf{X}}^{-1}\mathrm{Cov}(\mathbf{X}_{t+\tau}, Z_t),\\ \hat{a}_{\mathrm{FC}_\tau} &= \mu_Z - \widehat{\mathbf{W}}_{\mathrm{FC}_\tau}^\top \boldsymbol{\mu}.\end{aligned}$$

We hence have that

$$\begin{aligned}\min_{\substack{\mathbf{W}\in\mathbb{R}^N\\ a\in\mathbb{R}}} \mathrm{E}\left[\left((T_{-\tau}\mathbf{Z})_t - \mathbf{W}^\top U^F(\mathbf{Z})_t - a\right)^2\right] &= \mathrm{E}\left[\left(Z_{t+\tau} - \widehat{\mathbf{W}}_{\mathrm{MC}_\tau}^\top \mathbf{X}_t - \hat{a}_{\mathrm{MC}_\tau}\right)^2\right]\\ &= \mathrm{Var}(Z_{t+\tau}) - \mathrm{Cov}(\mathbf{X}_t, Z_{t+\tau})^\top \Gamma_{\mathbf{X}}^{-1}\mathrm{Cov}(\mathbf{X}_t, Z_{t+\tau})\end{aligned}$$

and

$$\begin{aligned}\min_{\substack{\mathbf{W}\in\mathbb{R}^N\\ a\in\mathbb{R}}} \mathrm{E}\left[\left(Z_t - \mathbf{W}^\top U^F(T_{-\tau}(\mathbf{Z}))_t - a\right)^2\right] &= \mathrm{E}\left[\left(Z_t - \widehat{\mathbf{W}}_{\mathrm{FC}_\tau}^\top \mathbf{X}_{t+\tau} - \hat{a}_{\mathrm{FC}_\tau}\right)^2\right]\\ &= \mathrm{Var}(Z_t) - \mathrm{Cov}(\mathbf{X}_{t+\tau}, Z_t)^\top \Gamma_{\mathbf{X}}^{-1}\mathrm{Cov}(\mathbf{X}_{t+\tau}, Z_t),\end{aligned}$$

which substituted in (3.1) and (3.2) and using the variance stationarity of $\mathbf{Z}$ yield (3.4), as required. ■

A.5. Proof of Lemma 3.3

In the proof we use the following elementary definition and linear algebraic fact: let $(V, \langle\cdot,\cdot\rangle_V)$ and $(W, \langle\cdot,\cdot\rangle_W)$ be two inner product spaces and let $f : V \longrightarrow W$ be a linear map between them. The dual map $f^* : W \longrightarrow V$ of f is defined by

$$\langle f^*(\mathbf{w}), \mathbf{v}\rangle_V = \langle \mathbf{w}, f(\mathbf{v})\rangle_W, \quad \text{for any} \quad \mathbf{v} \in V, \mathbf{w} \in W.$$

If the map f is injective, then the inner product $\langle\cdot,\cdot\rangle_W$ in W induces an inner product $\langle\cdot,\cdot\rangle_f$ in V via the equality

$$\langle \mathbf{v}_1, \mathbf{v}_2\rangle_f := \langle f(\mathbf{v}_1), f(\mathbf{v}_2)\rangle_W. \qquad \text{(A.3)}$$

It is easy to see that, in that case, the dual map with respect to the inner product (A.3) in V and $\langle\cdot,\cdot\rangle_W$ in W, satisfies that

$$f^*\circ f=\mathbb{I}_V, \tag{A.4}$$

and, in particular, $f^*: W\longrightarrow V$ is surjective.

We now proceed with the proof of the lemma. First of all, note that the echo state property hypothesis on F_2 and part **(ii)** of Proposition 2.3 imply that the filter U^{F_2} is well-defined and, moreover, for any $\mathbf{W}\in\mathbb{R}^{N_2}$ so is $U^{F_1}_{\mathbf{W}\circ f}$ and

$$U^{F_1}_{\mathbf{W}\circ f}=U^{F_2}_{\mathbf{W}}. \tag{A.5}$$

Let $\langle\cdot,\cdot\rangle_{\mathbb{R}^{N_2}}$ be the Euclidean inner product in $\mathbb{R}^{N_2}$, let $\langle\cdot,\cdot\rangle_{\mathbb{R}^{N_1}}$ be the inner product induced in $\mathbb{R}^{N_1}$ by $\langle\cdot,\cdot\rangle_{\mathbb{R}^{N_2}}$ and the injective map f using (A.3), and let f^* be the corresponding dual map. It is easy to see that the equality (A.5) can be rewritten using f^* as

$$U^{F_1}_{f^*(\mathbf{W})}=U^{F_2}_{\mathbf{W}},\quad \text{for any}\quad \mathbf{W}\in\mathbb{R}^{N_2}. \tag{A.6}$$

Let now $\tau\in\mathbb{Z}_-$ and let MC_τ be the τ-lag memory capacity of F_2 with respect to $\mathbf{Z}$. By definition and (A.6)

$$\begin{aligned}\mathrm{MC}_\tau &= 1-\frac{1}{\mathrm{Var}\,(Z_t)}\min_{\substack{\mathbf{W}\in\mathbb{R}^{N_2}\\ a\in\mathbb{R}}}\mathrm{E}\left[\left((T_{-\tau}\mathbf{Z})_t-U^{F_2}_{\mathbf{W}}(\mathbf{Z})_t-a\right)^2\right]\\ &=1-\frac{1}{\mathrm{Var}\,(Z_t)}\min_{\substack{\mathbf{W}\in\mathbb{R}^{N_2}\\ a\in\mathbb{R}}}\mathrm{E}\left[\left((T_{-\tau}\mathbf{Z})_t-U^{F_1}_{f^*(\mathbf{W})}(\mathbf{Z})_t-a\right)^2\right]\\ &=1-\frac{1}{\mathrm{Var}\,(Z_t)}\min_{\substack{\overline{\mathbf{W}}\in\mathbb{R}^{N_1}\\ a\in\mathbb{R}}}\mathrm{E}\left[\left((T_{-\tau}\mathbf{Z})_t-U^{F_1}_{\overline{\mathbf{W}}}(\mathbf{Z})_t-a\right)^2\right],\end{aligned}$$

which coincides with the τ-lag memory capacity of F_1. Notice that in the last equality we used the surjectivity of f^* which is a consequence of the injectivity of f (see (A.4)). A similar statement can be written for the forecasting capacities. ■

A.6. Proof of Theorem 3.4

First of all, since by hypothesis $\Gamma_{\mathbf{X}}$ is non-singular, Proposition 2.5 and the second part of Proposition 2.3 allow us to replace the system (2.1a)–(2.1b) in the definitions (3.1) and (3.2) by its standardized counterpart whose states $\widetilde{\mathbf{X}}_t$ are such that $\mathrm{E}[\widetilde{\mathbf{X}}_t]=\mathbf{0}$ and $\mathrm{Cov}(\widetilde{\mathbf{X}}_t,\widetilde{\mathbf{X}}_t)=\mathrm{E}\left[\widetilde{\mathbf{X}}_t\widetilde{\mathbf{X}}_t^\top\right]=\mathbb{I}_N$. More specifically, if we denote $\sigma^2:=\mathrm{Var}\,(Z_t)=\gamma(0)$, we can write

$$\begin{aligned}\mathrm{MC}_\tau &:= 1-\frac{1}{\sigma^2}\min_{\substack{\mathbf{W}\in\mathbb{R}^{N}\\ a\in\mathbb{R}}}\mathrm{E}\left[\left(Z_{t+\tau}-\mathbf{W}^\top\mathbf{X}_t-a\right)^2\right]\\ &=1-\frac{1}{\sigma^2}\min_{\substack{\widetilde{\mathbf{W}}\in\mathbb{R}^{N}\\ \widetilde{a}\in\mathbb{R}}}\mathrm{E}\left[\left(Z_{t+\tau}-\widetilde{\mathbf{W}}^\top\widetilde{\mathbf{X}}_t-\widetilde{a}\right)^2\right],\end{aligned}$$

which, using Lemma 3.2 and the fact that $\Gamma_{\widetilde{\mathbf{X}}}=\mathbb{I}_N$, can be rewritten as

$$\begin{aligned}\mathrm{MC}_\tau &= \frac{1}{\sigma^2}\mathrm{Cov}\left(Z_{t+\tau},\widetilde{\mathbf{X}}_t\right)\mathrm{Cov}\left(\widetilde{\mathbf{X}}_t,Z_{t+\tau}\right)\\ &=\frac{1}{\sigma^2}\sum_{i=1}^N\mathrm{E}\left[\widetilde{X}_t^i(Z_{t+\tau}-\mathrm{E}\left[Z_{t+\tau}\right])\right]^2.\end{aligned} \tag{A.7}$$

Analogously, it is easy to show that

$$\mathrm{FC}_\tau=\frac{1}{\sigma^2}\sum_{i=1}^N\mathrm{E}\left[\widetilde{X}^i_{t+\tau}(Z_t-\mathrm{E}\left[Z_t\right])\right]^2. \tag{A.8}$$

If we now define $\widetilde{Z}_t:=(Z_t-\mathrm{E}\left[Z_t\right])/\sigma\in L^2(\Omega,\mathbb{R})$, for any $t\in\mathbb{Z}_-$, it is clear that $\left\|\widetilde{Z}_t\right\|^2_{L^2}=\mathrm{E}\left[\widetilde{Z}_t^2\right]=1$. Moreover, the relation $\mathrm{Cov}(\widetilde{\mathbf{X}}_t,\widetilde{\mathbf{X}}_t)=\mathbb{I}_N$ implies that the components $\widetilde{X}_t^i\in L^2(\Omega,\mathbb{R})$, $i\in\{1,\ldots,N\}$, and that they form an orthonormal set, that is, $\left\langle\widetilde{X}_t^i,\widetilde{X}_t^j\right\rangle_{L^2}=\delta_{ij}$, where δ_{ij} stands for Kronecker's delta. It is actually the properties of the orthogonal projections onto the vector space generated by this orthonormal set that constitute the main technical tool in the proof and that will provide us with the capacity bounds that we are after. We now separately prove the three parts of the theorem.

(i) The fact that $\mathrm{MC}_\tau,\mathrm{FC}_\tau\geq 0$ is obvious from (A.7) and (A.8). Let now $\widetilde{S}_t=\mathrm{span}\left\{\widetilde{X}_t^1,\ldots,\widetilde{X}_t^N\right\}\subset L^2(\Omega,\mathbb{R})$ and let $\mathbb{P}_{\widetilde{S}_t}: L^2(\Omega,\mathbb{R})\longrightarrow\widetilde{S}_t$ be the corresponding orthogonal projection. Then,

$$\begin{aligned}1&=\left\|\widetilde{Z}_{t+\tau}\right\|^2_{L^2}\geq\left\|\mathbb{P}_{\widetilde{S}_t}\left(\widetilde{Z}_{t+\tau}\right)\right\|^2_{L^2}=\left\|\sum_{i=1}^N\widetilde{X}_t^i\left\langle\widetilde{X}_t^i,\widetilde{Z}_{t+\tau}\right\rangle_{L^2}\right\|^2_{L^2}\\ &=\sum_{i=1}^N\mathrm{E}\left[\widetilde{X}_t^i\widetilde{Z}_{t+\tau}\right]^2=\mathrm{MC}_\tau.\end{aligned}$$

The inequality $\mathrm{FC}_\tau\leq 1$ can be established analogously by considering the projection onto $\widetilde{S}_{t+\tau}$ of the vector $\widetilde{Z}_t$.

(ii) Define first, for any $L\in\mathbb{Z}_-$, the vector $\mathbf{Z}^L:=(Z_0-\mathrm{E}\left[Z_0\right], Z_{-1}-\mathrm{E}\left[Z_{-1}\right],\ldots,Z_L-\mathrm{E}\left[Z_L\right])$, and

$$\begin{aligned}\mathrm{MC}^L&:=\sum_{\tau=0}^L\mathrm{MC}_\tau=\sum_{\tau=0}^L\sum_{i=1}^N\mathrm{E}\left[\widetilde{X}_0^i\widetilde{Z}_\tau\right]^2,\\ \mathrm{FC}^L&:=\sum_{\tau=-1}^L\mathrm{FC}_\tau=\sum_{\tau=-1}^L\sum_{i=1}^N\mathrm{E}\left[\widetilde{X}_L^i\widetilde{Z}_{L-\tau}\right]^2.\end{aligned} \tag{A.9}$$

In these equalities we used (A.7) and (A.8) as well as the stationarity hypothesis. Now, the properties of the autocovariance function of a second-order stationary process guarantee that the matrix H^L is positive semidefinite (see [48, Theorem 1.5.1] and [51] for other properties) and since by hypothesis it is additionally invertible, we can associate to it a square root matrix $\left(H^L\right)^{1/2}$ that is also invertible. Hence, we define the random vector $\widehat{\mathbf{Z}}^L:=\left(H^L\right)^{-1/2}\mathbf{Z}^L$, whose components form an orthonormal set in $L^2(\Omega,\mathbb{R})$. Indeed, for any $i,j\in\{1,\ldots,-L+1\}$,

$$\begin{aligned}\langle\widehat{Z}_i^L,\widehat{Z}_j^L\rangle_{L^2}&=\mathrm{E}\left[\widehat{Z}_i^L\widehat{Z}_j^L\right]=\sum_{k,l=1}^{-L+1}\left(H^L\right)^{-1/2}_{ik}\left(H^L\right)^{-1/2}_{jl}\\ &\quad\times\mathrm{E}\left[(Z_{-k+1}-\mathrm{E}\left[Z_{-k+1}\right])(Z_{-l+1}-\mathrm{E}\left[Z_{-l+1}\right])\right]\\ &=\sum_{k,l=1}^{-L+1}\left(H^L\right)^{-1/2}_{ik}\left(H^L\right)^{-1/2}_{jl}\gamma(|k-l|)\\ &=\left(\left(H^L\right)^{-1/2}H^L\left(H^L\right)^{-1/2}\right)_{ij}=\delta_{ij}.\end{aligned} \tag{A.10}$$

Let $\widehat{S}_L=\mathrm{span}\left\{\widehat{Z}_1^L,\ldots,\widehat{Z}^L_{-L+1}\right\}\subset L^2(\Omega,\mathbb{R})$ and let $\mathbb{P}_{\widehat{S}_L}: L^2(\Omega,\mathbb{R})\longrightarrow\widehat{S}_L$ be the corresponding orthogonal projection. By (A.9) and using the definitions introduced earlier we have that

$$\begin{aligned}\mathrm{MC}^L&=\sum_{\tau=0}^L\sum_{i=1}^N\mathrm{E}\left[\widetilde{X}_0^i\widetilde{Z}_\tau\right]^2=\sum_{\tau=0}^L\sum_{i=1}^N\mathrm{E}\left[\widetilde{X}_0^i\frac{Z_\tau-\mathrm{E}\left[Z_\tau\right]}{\sigma}\right]^2\\ &=\frac{1}{\gamma(0)}\sum_{\tau=0}^L\sum_{i=1}^N\mathrm{E}\left[\widetilde{X}_0^i(\mathbf{Z}^L)_{-\tau+1}\right]^2\\ &=\frac{1}{\gamma(0)}\sum_{\tau=0}^L\sum_{i=1}^N\mathrm{E}\left[\widetilde{X}_0^i\left(\left(H^L\right)^{1/2}\widehat{\mathbf{Z}}^L\right)_{-\tau+1}\right]^2\\ &=\frac{1}{\gamma(0)}\sum_{i=1}^N\left\|\left(H^L\right)^{1/2}\mathrm{E}\left[\widetilde{X}_0^i\widehat{\mathbf{Z}}^L\right]\right\|^2.\end{aligned} \tag{A.11}$$

Analogously,

$$\begin{aligned}
\mathrm{FC}^L &= \sum_{\tau=-1}^{L}\sum_{i=1}^{N} \mathrm{E}\left[\widetilde{X}_L^i \widetilde{Z}_{L-\tau}\right]^2 \\
&= \frac{1}{\gamma(0)}\sum_{\tau=-1}^{L}\sum_{i=1}^{N} \mathrm{E}\left[\widetilde{X}_L^i \left(\left(H^L\right)^{1/2}\widehat{\mathbf{Z}}^L\right)_{-L+\tau+1}\right]^2 \\
&\leq \frac{1}{\gamma(0)}\sum_{i=1}^{N}\left\|\left(H^L\right)^{1/2}\mathrm{E}\left[\widetilde{X}_L^i\widehat{\mathbf{Z}}^L\right]\right\|^2 . \qquad \text{(A.12)}
\end{aligned}$$

Now, by (A.10) and (A.11) we can write that

$$\begin{aligned}
\mathrm{MC}^L &\leq \frac{1}{\gamma(0)}\sum_{i=1}^{N}\left|\!\left|\!\left|\left(H^L\right)^{1/2}\right|\!\right|\!\right|^2 \left\|\mathrm{E}\left[\widetilde{X}_0^i\widehat{\mathbf{Z}}^L\right]\right\|^2 \\
&= \frac{1}{\gamma(0)}\sum_{i=1}^{N}\left|\!\left|\!\left|\left(H^L\right)^{1/2}\right|\!\right|\!\right|^2 \left\|\mathbb{P}_{\widehat{S}_L}\left(\widetilde{X}_0^i\right)\right\|_{L^2}^2 \\
&\leq \frac{1}{\gamma(0)}\sum_{i=1}^{N}\left|\!\left|\!\left|\left(H^L\right)^{1/2}\right|\!\right|\!\right|^2 \left\|\widetilde{X}_0^i\right\|_{L^2}^2 \\
&\leq \frac{N}{\gamma(0)}\left|\lambda_{\max}(H^L)\right| = \frac{N}{\gamma(0)}\rho(H^L). \qquad \text{(A.13)}
\end{aligned}$$

An identical inequality can be shown for FC^L using (A.12), which proves the first inequality in (3.6). The second inequality can be obtained by bounding $\rho(H^L)$ using Gershgorin's Disks Theorem (see [52, Theorem 6.1.1 and Corollary 6.1.5]). Indeed, due to this result:

$$\rho(H^L) \leq \gamma(0) + \max_{i\in\{1,\ldots,-L+1\}}\left\{\sum_{\substack{j\in\{1,\ldots,-L+1\}\\ j\neq i}}\left|H_{ij}^L\right|\right\} \leq \gamma(0) + 2\sum_{i=1}^{-L}|\gamma(i)|. \qquad \text{(A.14)}$$

(iii) The inequalities in (3.8) are a consequence of considering H as the infinite symmetric Toeplitz matrix associated to the bi-infinite sequence of autocovariances $\{\gamma(j)\}_{j\in\mathbb{Z}}$ of $\mathbf{Z}$. First, when the autocovariance function is absolutely summable then (3.7) determines the spectral density of $\mathbf{Z}$ by [48, Corollary 4.3.2]. Second, by [53, Lemma 6, page 194], the spectrum of H is bounded above by the maximum of the function $2\pi f$, which, using (3.6) implies that $\mathrm{C} \leq \frac{2\pi N}{\gamma(0)}M_f$. The last inequality is a consequence of

$$|f(\lambda)| \leq \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}|\gamma(j)| = \frac{1}{2\pi}\left(\gamma(0) + 2\sum_{j=1}^{\infty}|\gamma(j)|\right) < +\infty,$$

for any $\lambda \in [-\pi,\pi]$. ■

A.7. Proof of Corollary 3.5

The first inequality is a straightforward consequence of (3.6) and the fact that for independent inputs $\gamma(h) = 0$, for all $h \neq 0$. The second one can be easily obtained from (A.8). Indeed, by the causality and the time-invariance [34, Proposition 2.1] of any filter induced by a state-space system of the type (2.1a)–(2.1b), for any $\tau \leq -1$, the random variables $\widetilde{X}_{t+\tau}^i$ and Z_t in (A.8) are independent, and hence

$$\begin{aligned}
\mathrm{FC}_\tau &= \frac{1}{\sigma^2}\sum_{i=1}^{N}\mathrm{E}\left[\widetilde{X}_{t+\tau}^i(Z_t - \mathrm{E}[Z_t])\right]^2 \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{N}\mathrm{E}\left[\widetilde{X}_{t+\tau}^i\right]^2 \mathrm{E}\left[Z_t - \mathrm{E}[Z_t]\right]^2 = 0. \quad ■
\end{aligned}$$

A.8. Proof of Proposition 4.1

First of all, recall that the matrix norm $|\!|\!|\cdot|\!|\!|$ induced by the Euclidean norm $\|\cdot\|$ in $\mathbb{R}^N$ is defined as

$$|\!|\!|A|\!|\!| = \sup_{\mathbf{x}\in\mathbb{R}^N,\mathbf{x}\neq\mathbf{0}}\left\{\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}\right\} = \sigma_{\max}(A).$$

It follows from this definition that the condition $\sigma_{\max}(A) < 1$ implies that the state map $F(\mathbf{x}, z)$ in (4.1) is a contraction on the first entry. Indeed, for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^N$, $z \in \mathbb{R}$, we have

$$\|F(\mathbf{x}_1, z) - F(\mathbf{x}_2, z)\| = \|A(\mathbf{x}_1 - \mathbf{x}_2)\| \leq |\!|\!|A|\!|\!|\,\|\mathbf{x}_1 - \mathbf{x}_2\| < \|\mathbf{x}_1 - \mathbf{x}_2\|.$$

Second, as the input process takes values on a compact set, there exists a compact subset $D_N \subset \mathbb{R}^N$ (see [46, Remark 2]) such that the restriction $F : D_N \times D \longrightarrow D_N$ satisfies the hypotheses of Proposition 2.1 and of Corollary 2.4. This implies that the system associated to F has the echo state property, as well as the stationarity of the filter output $\mathbf{X} = U^{A,\mathbf{C}}(\mathbf{Z})$ and of the joint processes in the statement. We recall that, in this case,

$$\mathbf{X}_t = U^{A,\mathbf{C}}(\mathbf{Z})_t = \sum_{j=0}^{\infty}A^j\mathbf{C}Z_{t-j}, \quad \text{for any} \quad t \in \mathbb{Z}_-.$$

We now show that the output process is also square-integrable and hence covariance stationary. Indeed, let $\mathbf{X}_t^n := \sum_{j=0}^{n}A^j\mathbf{C}Z_{t-j}$, $n \in \mathbb{N}$. Given that by hypothesis D is compact, there exists $M > 0$ such that $D \subset [-M, M]$ and hence

$$\left\|\mathbf{X}_t^n\right\| = \left\|\sum_{j=0}^{n}A^j\mathbf{C}Z_{t-j}\right\| \leq M\,\|\mathbf{C}\|\sum_{j=0}^{n}|\!|\!|A|\!|\!|^j \leq \frac{M\,\|\mathbf{C}\|}{1-\sigma_{\max}(A)}.$$

Now the Bounded Convergence Theorem guarantees that

$$\|\mathbf{X}_t\|_{L^2} = \mathrm{E}\left[\|\mathbf{X}_t\|^2\right]^{\frac{1}{2}} = \lim_{n\to\infty}\mathrm{E}\left[\left\|\mathbf{X}_t^n\right\|^2\right]^{\frac{1}{2}} \leq \frac{M\,\|\mathbf{C}\|}{1-\sigma_{\max}(A)} < \infty.$$

The Cauchy–Schwarz inequality implies that the components of $\Gamma_{\mathbf{X}} = \mathrm{E}\left[\mathbf{X}_t\mathbf{X}_t^\top\right]$ are also finite and hence, using the notation introduced in Proposition 2.5 and the invertibility hypothesis on $\Gamma_{\mathbf{X}}$, the standardized states $\widetilde{\mathbf{X}}_t$ are given by

$$\widetilde{\mathbf{X}}_t = \Gamma_{\mathbf{X}}^{-1/2}\mathbf{X}_t = \Gamma_{\mathbf{X}}^{-1/2}\sum_{j=0}^{\infty}A^j\mathbf{C}Z_{t-j}. \qquad \text{(A.15)}$$

Additionally, when the autocovariance function of the input is absolutely summable, then the spectral density f of $\mathbf{Z}$ defined in (3.7) belongs to the so-called Wiener class and, moreover, if the hypothesis on it in the statement is satisfied, then the two matrices H (semi-infinite) and $\overline{H}$ (doubly infinite) are invertible (see [53, Theorem 11]).

(i) Let $\overline{\mathbf{Z}} := (Z_0, Z_1, \ldots)$ and let $\widehat{\mathbf{Z}} = H^{-1/2}\overline{\mathbf{Z}}$. An argument similar to (A.10) shows that $\langle\widehat{Z}_i, \widehat{Z}_j\rangle_{L^2} = \delta_{ij}$. Moreover, (A.15) implies that

$$\begin{aligned}
\widetilde{\mathbf{X}}_0 &= \Gamma_{\mathbf{X}}^{-1/2}\sum_{k=0}^{\infty}A^k\mathbf{C}Z_{-k} = \Gamma_{\mathbf{X}}^{-1/2}\sum_{k=0}^{\infty}A^k\mathbf{C}\left(H^{1/2}\widehat{\mathbf{Z}}\right)_{k+1} \\
&= \Gamma_{\mathbf{X}}^{-1/2}\sum_{k=0}^{\infty}\sum_{j=1}^{\infty}A^k\mathbf{C}H_{k+1,j}^{1/2}\widehat{Z}_j,
\end{aligned}$$

which, using the definition in (4.2) can be rewritten componentwise as

$$\widetilde{X}_0^i = \sum_{j=1}^{\infty}B_i^j\widehat{Z}_j, \qquad i \in \{1,\ldots,N\}. \qquad \text{(A.16)}$$

The scalars B_i^j are defined in (4.2) and by (A.16) coincide with the (unique) coefficients that determine the expansion of $\widetilde{X}_0^i$ on the orthonormal basis $\left\{\widehat{Z}_t\right\}_{t\in\mathbb{Z}_-}$. This implies, in particular, that $B_i^j = \langle \widetilde{X}_0^i, \widehat{Z}_j\rangle_{L^2}$ and, by the Cauchy–Schwarz inequality, that all these coefficients are finite.

We now show that the L^2-orthonormality of the components $\widetilde{X}_0^i$ of $\widetilde{\mathbf{X}}_0$ implies the ℓ^2-orthonormality of the vectors $\{\mathbf{B}_1, \ldots, \mathbf{B}_N\} \in \ell_+^2(\mathbb{R})$ whose components we just showed are finite. Indeed, for any $i, j \in \{1, \ldots, N\}$, the equality (A.16) and the Parseval identity imply that

$$\delta_{ij} = \langle \widetilde{X}_0^i, \widetilde{X}_0^j\rangle_{L^2} = \sum_{k=1}^{\infty} B_i^k B_j^k = \langle \mathbf{B}_i, \mathbf{B}_j\rangle_{\ell_+^2}. \tag{A.17}$$

We now prove (4.3). First, taking the limit $L \to \infty$ in (A.11) we write,

$$\mathrm{MC} = \frac{1}{\gamma(0)}\sum_{i=1}^{N}\left\|H^{1/2}\mathrm{E}\left[\widetilde{X}_0^i\widehat{\mathbf{Z}}\right]\right\|^2 = \frac{1}{\gamma(0)}\sum_{i=1}^{N}\left\|H^{1/2}\sum_{j=1}^{\infty}B_i^j\mathrm{E}\left[\widehat{Z}_j\widehat{\mathbf{Z}}\right]\right\|^2$$
$$= \frac{1}{\gamma(0)}\sum_{i=1}^{N}\left\|H^{1/2}\mathbf{B}_i\right\|_{\ell^2}^2 = \frac{1}{\gamma(0)}\sum_{i=1}^{N}\langle \mathbf{B}_i, H\mathbf{B}_i\rangle_{\ell^2}.$$

(ii) Let now $\overline{\mathbf{Z}} := (\ldots, Z_{-1}, Z_0, Z_1, \ldots)$ and let $\widehat{\mathbf{Z}} = \overline{H}^{-1/2}\overline{\mathbf{Z}}$. An argument similar to (A.10) shows that $\langle \widehat{Z}_i, \widehat{Z}_j\rangle_{L^2} = \delta_{ij}$ for any $i, j \in \mathbb{Z}$. Also, in this case, (A.15) implies that

$$\widetilde{\mathbf{X}}_0 = \Gamma_{\mathbf{X}}^{-1/2}\sum_{k=0}^{\infty}\sum_{j=-\infty}^{\infty}A^k\mathbf{C}\overline{H}_{-k,j}^{1/2}\widehat{Z}_j,$$

which, using the definition in (4.4) can be rewritten componentwise as

$$\widetilde{X}_0^i = \sum_{j=-\infty}^{\infty}B_i^j\widehat{Z}_j, \qquad i \in \{1, \ldots, N\}.$$

As in (A.17), the L^2-orthonormality of the components $\widetilde{X}_0^i$ of $\widetilde{\mathbf{X}}_0$ implies the ℓ^2-orthonormality of the vectors $\{\mathbf{B}_1, \ldots, \mathbf{B}_N\} \in \ell^2(\mathbb{R})$. Finally, in order to establish (4.5), we use the stationarity hypothesis to rewrite the expression of the forecasting capacity in (A.8) as

$$\mathrm{FC} = \frac{1}{\gamma(0)}\sum_{\tau=1}^{\infty}\sum_{i=1}^{N}\mathrm{E}\left[\widetilde{X}_0^iZ_\tau\right]^2 = \frac{1}{\gamma(0)}\sum_{\tau=1}^{\infty}\sum_{i=1}^{N}\mathrm{E}\left[\sum_{j=-\infty}^{\infty}B_i^j\widehat{Z}_j\left(\overline{H}^{1/2}\widehat{\mathbf{Z}}\right)_\tau\right]^2$$
$$= \frac{1}{\gamma(0)}\sum_{\tau=1}^{\infty}\sum_{i=1}^{N}\left(\sum_{j,l=-\infty}^{\infty}B_i^j\overline{H}_{\tau,l}^{1/2}\mathrm{E}\left[\widehat{Z}_j\widehat{Z}_l\right]\right)^2$$
$$= \frac{1}{\gamma(0)}\sum_{\tau=1}^{\infty}\sum_{i=1}^{N}\left(\sum_{j=-\infty}^{\infty}B_i^j\overline{H}_{\tau,j}^{1/2}\right)^2$$
$$= \frac{1}{\gamma(0)}\sum_{\tau=1}^{\infty}\sum_{i=1}^{N}\left(\overline{H}^{1/2}\mathbf{B}_i\right)_\tau^2$$
$$= \frac{1}{\gamma(0)}\sum_{i=1}^{N}\left\|\mathbb{P}_{\mathbb{Z}^+}\left(\overline{H}^{1/2}\mathbf{B}_i\right)\right\|_{\ell^2}^2. \quad \blacksquare$$

A.9. Proof of Corollary 4.2

The first equality in (4.6) is a straightforward consequence of (4.3) and of the fact that for white noise inputs $H = \gamma(0)\mathbb{I}_{\ell_+^2(\mathbb{R})}$. The equality $\mathrm{FC} = 0$ follows from the fact that $\overline{H} = \gamma(0)\mathbb{I}_{\ell^2(\mathbb{R})}$ and that $B_i^j = 0$, for any $i \in \{1, \ldots, N\}$ and any $j \in \mathbb{Z}^+$, by (4.4). Then, by (4.5),

$$\mathrm{FC} = \frac{1}{\gamma(0)}\sum_{i=1}^{N}\left\|\mathbb{P}_{\mathbb{Z}^+}\left(\overline{H}^{1/2}\mathbf{B}_i\right)\right\|_{\ell^2}^2 = \sum_{i=1}^{N}\left\|\mathbb{P}_{\mathbb{Z}^+}(\mathbf{B}_i)\right\|_{\ell^2}^2 = 0. \quad \blacksquare$$

A.10. Proof of Proposition 4.3

(i) Using the notation introduced in the statement notice that:

$$A\mathbf{C} = \sum_{i=1}^{N}c_i\lambda_i\mathbf{v}_i,\ A^2\mathbf{C} = \sum_{i=1}^{N}c_i\lambda_i^2\mathbf{v}_i,\ \ldots,\ A^N\mathbf{C} = \sum_{i=1}^{N}c_i\lambda_i^N\mathbf{v}_i. \tag{A.18}$$

Since by hypothesis $|||A||| = \sigma_{\max}(A) < 1$, the spectral radius $\rho(A)$ of A satisfies that $\rho(A) \le \sigma_{\max}(A) < 1$, and hence:

$$B := \sum_{j=0}^{\infty}A^j\mathbf{C}\mathbf{C}^\top\left(A^j\right)^\top = \sum_{k=0}^{\infty}\sum_{i,j=1}^{N}\lambda_i^k\lambda_j^kc_ic_j\mathbf{v}_i\mathbf{v}_j^\top = \sum_{i,j=1}^{N}\frac{c_ic_j}{1-\lambda_i\lambda_j}\mathbf{v}_i\mathbf{v}_j^\top, \tag{A.19}$$

which shows that in the matrix basis $\left\{\mathbf{v}_i\mathbf{v}_j^\top\right\}_{i,j\in\{1,\ldots,N\}}$, the matrix B has components $\overline{B}_{ij} := \frac{c_ic_j}{1-\lambda_i\lambda_j}$ or, equivalently,

$$\overline{B} := \overline{\mathbf{C}}\overline{\mathbf{C}}^\top \odot D, \quad \text{with} \quad \overline{\mathbf{C}} = (c_1, \ldots, c_N)^\top,$$
$$\text{and } D \text{ defined by} \quad D_{ij} := \frac{1}{1-\lambda_i\lambda_j}, \tag{A.20}$$

for any $i, j \in \{1, \ldots, N\}$; the symbol $\odot$ stands for componentwise matrix multiplication (Hadamard product). Let $P \in \mathbb{M}_N$ be the invertible change-of-basis matrix between $\{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$ and the canonical basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_N\}$, that is, for any $i \in \{1, \ldots, N\}$, we have that $\mathbf{v}_i = \sum_{k=1}^{N}P_{ik}\mathbf{e}_k$. It is easy to see that $B = P^\top\overline{B}P$ and hence the invertibility of B (and hence of $\Gamma_{\mathbf{X}}$) is equivalent to the invertibility of $\overline{B}$ in (A.20), which we now characterize.

In order to provide an alternative expression for $\overline{B}$, recall that for any two vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^N$, we can write $\mathbf{v} \odot \mathbf{w} = \mathrm{diag}(\mathbf{v})\mathbf{w}$, where $\mathrm{diag}(\mathbf{v}) \in \mathbb{M}_N$ is the diagonal matrix that has the entries of the vector $\mathbf{v}$ in the diagonal. Using this fact and the Hadamard product trace property (see [54, Lemma 5.1.4, page 305]) we have that

$$\langle\mathbf{v}, \overline{B}\mathbf{w}\rangle = \left\langle\mathbf{v}, (\overline{\mathbf{C}}\overline{\mathbf{C}}^\top \odot D)\mathbf{w}\right\rangle = \mathrm{trace}\left(\mathbf{v}^\top(\overline{\mathbf{C}}\overline{\mathbf{C}}^\top \odot D)\mathbf{w}\right)$$
$$= \mathrm{trace}\left((\overline{\mathbf{C}}\overline{\mathbf{C}}^\top \odot D)\mathbf{w}\mathbf{v}^\top\right)$$
$$= \mathrm{trace}\left((\overline{\mathbf{C}}\overline{\mathbf{C}}^\top \odot \mathbf{v}\mathbf{w}^\top)D^\top\right) = \mathrm{trace}\left((\overline{\mathbf{C}} \odot \mathbf{v})(\overline{\mathbf{C}} \odot \mathbf{w})^\top D^\top\right)$$
$$= \mathrm{trace}\left(\mathrm{diag}(\overline{\mathbf{C}})\mathbf{v}\mathbf{w}^\top\mathrm{diag}(\overline{\mathbf{C}})D^\top\right) = \left\langle\mathbf{v}, \mathrm{diag}(\overline{\mathbf{C}})D\mathrm{diag}(\overline{\mathbf{C}})\mathbf{w}\right\rangle.$$

Since $\mathbf{v}, \mathbf{w} \in \mathbb{R}^N$ are arbitrary, this equality allows us to conclude that $\overline{B} = \mathrm{diag}(\overline{\mathbf{C}})D\mathrm{diag}(\overline{\mathbf{C}})$ and hence $\overline{B}$ is invertible if and only if both $\mathrm{diag}(\overline{\mathbf{C}})$ and D are. The regularity of $\mathrm{diag}(\overline{\mathbf{C}})$ is equivalent to requiring that all the entries of the vector $\overline{\mathbf{C}}$ are non-zero. Regarding D, it can be shown by induction on the matrix dimension N, that

$$\det(D) = (-1)^N\frac{\prod_{i<j=1}^{N}\left(\lambda_i-\lambda_j\right)^2}{\prod_{i<j=1}^{N}\left(\lambda_i\lambda_j-1\right)^2\prod_{i=1}^{N}(\lambda_i^2-1)}.$$

Consequently, D is invertible if and only if $\det(D) \neq 0$, which is equivalent to all the elements in the spectrum $\sigma(A)$ being distinct.

(ii) The condition on the vectors $\left\{A\mathbf{C}, A^2\mathbf{C}, \ldots, A^N\mathbf{C}\right\}$ forming a basis of $\mathbb{R}^N$ is equivalent to the invertibility of the matrix $R(A, \mathbf{C}) := \left(A\mathbf{C}|A^2\mathbf{C}|\cdots|A^N\mathbf{C}\right)$. It is easy to see using (A.18) that

$$\widehat{R(A, \mathbf{C})} = P^\top\overline{R(A, \mathbf{C})}, \tag{A.21}$$

where P is the invertible change-of-basis matrix in the previous point and $\overline{R(A, \mathbf{C})}$ is given by

$$\overline{R(A,\mathbf{C})} := \begin{pmatrix} c_1\lambda_1 & c_1\lambda_1^2 & \cdots & c_1\lambda_1^N \\ c_2\lambda_2 & c_2\lambda_2^2 & \cdots & c_2\lambda_2^N \\ \vdots & \vdots & \ddots & \vdots \\ c_N\lambda_N & c_N\lambda_N^2 & \cdots & c_N\lambda_N^N \end{pmatrix}. \tag{A.22}$$

Indeed, for any $i, j \in \{1, \dots, N\}$,

$$\begin{aligned}\widehat{R(A,\mathbf{C})}_{ij} &= \left(A^j\mathbf{C}\right)_i = \left(\sum_{k=1}^N c_k A^j \mathbf{v}_k\right)_i = \left(\sum_{k=1}^N c_k \lambda_k^j \mathbf{v}_k\right)_i = \left(\sum_{k,l=1}^N c_k \lambda_k^j P_{kl}\mathbf{e}_l\right)_i \\ &= \left(\sum_{k,l=1}^N \overline{R(A,\mathbf{C})}_{kj} P_{kl}\mathbf{e}_l\right)_i = \left(\sum_{l=1}^N \left(P^\top \overline{R(A,\mathbf{C})}\right)_{lj}\mathbf{e}_l\right)_i \\ &= \sum_{l=1}^N \left(P^\top \overline{R(A,\mathbf{C})}\right)_{lj} \mathbf{e}_i^\top \mathbf{e}_l = \left(P^\top \overline{R(A,\mathbf{C})}\right)_{ij},\end{aligned}$$

which proves (A.21). Now, using induction on the matrix dimension N, it can be shown that

$$\det(\overline{R(A,\mathbf{C})}) = \prod_{i=1}^N c_i\lambda_i \prod_{i<j=1}^N (\lambda_i - \lambda_j).$$

The invertibility of $\overline{R(A,\mathbf{C})}$ (or, equivalently, the invertibility of $\widehat{R(A,\mathbf{C})}$) is equivalent to all the coefficients c_i and all the eigenvalues λ_i being non-zero (so that $\prod_{i=1}^N c_i\lambda_i$ is non-zero) and all the elements in $\sigma(A)$ being distinct (so that $\prod_{i<j=1}^N(\lambda_i - \lambda_j)$ is non-zero).

(iii) The Kalman controllability condition on the vectors $\left\{\mathbf{C}, A\mathbf{C}, \dots, A^{N-1}\mathbf{C}\right\}$ forming a basis of $\mathbb{R}^N$ is equivalent to the invertibility of the ***controllability*** or ***reachability*** matrix $R(A,\mathbf{C}) := \left(\mathbf{C}|A\mathbf{C}|\cdots|A^{N-1}\mathbf{C}\right)$ (see [30] for this terminology). Following the same strategy that we used to prove (A.21), it is easy to see that $R(A,\mathbf{C}) = P^\top \widetilde{R(A,\mathbf{C})}$, where

$$\widetilde{R(A,\mathbf{C})} := \begin{pmatrix} c_1 & c_1\lambda_1 & \cdots & c_1\lambda_1^{N-1} \\ c_2 & c_2\lambda_2 & \cdots & c_2\lambda_2^{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ c_N & c_N\lambda_N & \cdots & c_N\lambda_N^{N-1} \end{pmatrix}.$$

The matrix $\widetilde{R(A,\mathbf{C})}$ has the same rank as $\overline{R(A,\mathbf{C})}$ since it can be obtained from $\overline{R(A,\mathbf{C})}$ via elementary matrix operations, namely, by dividing each row i of $\overline{R(A,\mathbf{C})}$ by the corresponding eigenvalue λ_i which is by hypothesis non-zero. ■

Remark A.1. The matrix B in (A.19), which is instrumental for the proof of part **(i)** of the proposition, is related to the objects introduced in [49], in particular to the positive semi-definite symmetric matrix Q corresponding to the temporal kernel associated to the linear dynamical system and defined as $Q_{i,j} = \mathbf{C}^\top (A^{j-1})^\top A^{i-1}\mathbf{C}$, $i, j \in \{1, \dots, \tau\}$, $\tau \in \mathbb{N}$. More specifically, if one defines B_τ as the covariance matrix of the states process associated to the truncated solution of the linear state system (4.1) with white noise as inputs, that is, $B_\tau := \sum_{j=0}^\tau \mathbf{C}^\top \left(A^j\right)^\top A^j\mathbf{C}$, $\tau \in \mathbb{N}$, then it is easy to see that $\mathrm{trace}(B_\tau) = \sum_{j=1}^\tau \mathbf{C}^\top (A^j)^\top A^j \mathbf{C} = \sum_{j=1}^\tau Q_{j+1,j+1} = \mathrm{trace}(Q^d)$ with $Q^d \in \mathbb{S}_\tau$ the diagonal matrix with the same elements on the main diagonal as Q. The results in [49] provide bounds for the elements $Q_{i,j}$, $i, j \in \{1, \dots, \tau\}$, whenever the state map is constructed with a randomly generated connectivity matrix A and input matrix $\mathbf{C}$, for different choices of architectures and distributions. Obviously, the analysis of the diagonal entries of Q for those situations is valid for the diagonal elements of B_τ and hence, since $\Gamma_{\mathbf{X}} = \gamma(0)B$, for large τ also illustrates the behavior of the variances of the states process of linear state systems.

A.11. Proof of Theorem 4.4

(i) We first show that $\ker R(A,\mathbf{C})^\top \subset \ker \Gamma_{\mathbf{X}}$. Let $\mathbf{v} \in \ker R(A,\mathbf{C})^\top$. This implies that

$$\mathbf{C}^\top (A^j)^\top \mathbf{v} = \mathbf{0}, \text{ for all } j \in \{0, \dots, N-1\}. \tag{A.23}$$

Now, by the Hamilton–Cayley Theorem [52, Theorem 2.4.3.2], for any $j \geq N$, there exist constants $\left\{\beta_0^j, \dots, \beta_{N-1}^j\right\}$ such that $A^j = \sum_{i=0}^{N-1} \beta_i^j A^i$ which, together with (A.23), implies that equality holds for all $j \in \mathbb{N}$. Recall now that by Proposition 4.3 part **(i)**, $\Gamma_{\mathbf{X}} = \gamma(0)\sum_{j=0}^\infty A^j \mathbf{C}\mathbf{C}^\top \left(A^j\right)^\top$, and hence we can conclude that $\Gamma_{\mathbf{X}}(\mathbf{v}) = \gamma(0)\sum_{j=0}^\infty A^j\mathbf{C}\mathbf{C}^\top \left(A^j\right)^\top \mathbf{v} = \mathbf{0}$, that is, $\mathbf{v} \in \ker \Gamma_{\mathbf{X}}$. Conversely, if $\mathbf{v} \in \ker \Gamma_{\mathbf{X}}$, we have that $0 = \langle \mathbf{v}, \Gamma_{\mathbf{X}}(\mathbf{v})\rangle = \gamma(0)\sum_{j=0}^\infty \|\mathbf{C}^\top \left(A^j\right)^\top \mathbf{v}\|^2$, which implies that $\mathbf{C}^\top \left(A^j\right)^\top \mathbf{v} = \mathbf{0}$, necessarily, for any $j \in \mathbb{N}$ and hence $\mathbf{v} \in \ker R(A,\mathbf{C})^\top$.

(ii) The system associated to $\overline{F}$ has the echo state property because for any $\overline{\mathbf{x}} \in X$,

$$\left\|\overline{A}\overline{\mathbf{x}}\right\|^2 = \left\|\pi_X A i_X(\overline{\mathbf{x}})\right\|^2 \leq \left\|A i_X(\overline{\mathbf{x}})\right\|^2,$$

which implies that $|||\overline{A}||| \leq |||A||| = \sigma_{\max}(A) < 1$.

(iii) We first show that for any $\overline{\mathbf{x}} \in X$ and $z \in D$ we have that $A i_X \overline{\mathbf{x}} + \mathbf{C}z \in X$. Indeed, as $\overline{\mathbf{x}} \in X$, there exist constants $\{\alpha_0, \dots, \alpha_{N-1}\}$ such that $\overline{\mathbf{x}} = \sum_{i=0}^{N-1} \alpha_i A^i \mathbf{C}$ and hence

$$\begin{aligned} A i_X \overline{\mathbf{x}} + \mathbf{C}z &= \sum_{i=0}^{N-1} \alpha_i A^{i+1}\mathbf{C} + \mathbf{C}z = \mathbf{C}z + \sum_{i=1}^{N-1} \alpha_{i-1}A^i\mathbf{C} + \alpha_{N-1}A^N\mathbf{C} \\ &= \mathbf{C}z + \sum_{i=1}^{N-1}\alpha_{i-1}A^i\mathbf{C} + \alpha_{N-1}\sum_{j=0}^{N-1}\beta_j^N A^j\mathbf{C} \in X, \end{aligned} \tag{A.24}$$

where the constants $\left\{\beta_0^N, \dots, \beta_{N-1}^N\right\}$ satisfy that $A^N = \sum_{i=0}^{N-1}\beta_i^N A^i$ and, as above, are a byproduct of the Hamilton-Cayley Theorem [52, Theorem 2.4.3.2]. We now show that i_X is a system equivariant map between $\overline{F}$ and F. For any $\overline{\mathbf{x}} \in X$ and $z \in D$,

$$i_X\left(\overline{F}(\overline{\mathbf{x}}, z)\right) = i_X\pi_X\left(A i_X(\overline{\mathbf{x}}) + \mathbf{C}z\right) = A i_X(\overline{\mathbf{x}}) + \mathbf{C}z = F(i_X(\overline{\mathbf{x}}), z),$$

where the second equality holds because $i_X \circ \pi_X|_X = \mathbb{I}_X$ and, by (A.24), $A i_X \overline{\mathbf{x}} + \mathbf{C}z \in X$.

(iv) We prove the identity $\mathrm{rank}\, R(\overline{A}, \overline{\mathbf{C}}) = r$ by showing that

$$i_X\left(\mathrm{span}\left\{\overline{\mathbf{C}}, \overline{A}\overline{\mathbf{C}}, \dots, \overline{A}^{r-1}\overline{\mathbf{C}}\right\}\right) = X.$$

First, using that $i_X \circ \pi_X|_X = \mathbb{I}_X$ and that by the Hamilton–Cayley Theorem $A^j\mathbf{C} \in X$, for all $j \in \mathbb{N}$, it is easy to conclude that for any $\mathbf{v} = \sum_{i=1}^r \alpha_i \overline{A}^{i-1}\overline{\mathbf{C}}$ we have that

$$i_X(\mathbf{v}) = i_X\left(\sum_{i=1}^r \alpha_i \overline{A}^{i-1}\overline{\mathbf{C}}\right) = \sum_{i=1}^r \alpha_i A^{i-1}\mathbf{C},$$

which shows that $i_X\left(\mathrm{span}\left\{\overline{\mathbf{C}}, \overline{A}\overline{\mathbf{C}}, \dots, \overline{A}^{r-1}\overline{\mathbf{C}}\right\}\right) \subset X$. Conversely, let $\mathbf{v} = \sum_{i=1}^N \alpha_i A^{i-1}\mathbf{C} \in X$. Using again that $i_X \circ \pi_X|_X = \mathbb{I}_X$, we can write that,

$$\begin{aligned}\mathbf{v} = \sum_{i=1}^N \alpha_i A^{i-1}\mathbf{C} &= \sum_{i=1}^N \alpha_i \underbrace{(i_X\pi_X A)\cdots(i_X\pi_X A)}_{i\,-\,1\text{-times}}(i_X\pi_X\mathbf{C}) \\ &= i_X\left(\sum_{i=1}^N \alpha_i \overline{A}^{i-1}\overline{\mathbf{C}}\right)\end{aligned}$$

$$= i_X \left(\sum_{i=1}^{r} \alpha_i \overline{A}^{i-1} \overline{\mathbf{C}} + \sum_{j=r+1}^{N} \sum_{k_j=1}^{r} \alpha_j \beta_l^j \overline{A}^{l-1} \overline{\mathbf{C}} \right),$$

which clearly belongs to $i_X \left(\operatorname{span} \left\{ \overline{\mathbf{C}}, \overline{A\mathbf{C}}, \ldots, \overline{A}^{r-1} \overline{\mathbf{C}} \right\} \right)$. The constants β_l^j are obtained, again, by using the Hamilton–Cayley Theorem. Finally, the invertibility of $\Gamma_{\overline{\mathbf{X}}}$ is a consequence of Proposition 4.3 and the hypothesis that the elements of the spectrum $\sigma(\overline{A})$ are non-zero.

(v) The statement in part **(iii)** that we just proved and Lemma 3.3 imply that the memory MC and forecasting FC capacities of F with respect to $\mathbf{Z}$ coincide with those of $\overline{F}$ with respect to $\mathbf{Z}$. Now, the statement in part **(iv)** and Corollary 4.2 imply that those capacities coincide with r and 0, respectively. Finally, as $r = \operatorname{rank} R(A, \mathbf{C})$, the claim follows. ■

References

[1] H. Jaeger, Short Term Memory in Echo State Networks, Vol. 152, Technical Report, Fraunhofer Institute for Autonomous Intelligent Systems, 2002.
[2] M.B. Matthews, On the Uniform Approximation of Nonlinear Discrete-Time Fading-Memory Systems using Neural Network Models (Ph.D. thesis), ETH Zürich, 1992.
[3] M. Matthews, G. Moschytz, The identification of nonlinear discrete-time fading-memory systems using neural network models, IEEE Trans. Circuits Syst. II 41 (11) (1994) 740–751.
[4] H. Jaeger, H. Haas, Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication, Science 304 (5667) (2004) 78–80.
[5] O. White, D. Lee, H. Sompolinsky, Short-term memory in orthogonal neural networks, Phys. Rev. Lett. 92 (14) (2004) 148102.
[6] S. Ganguli, D. Huh, H. Sompolinsky, Memory traces in dynamical systems, Proc. Natl. Acad. Sci. USA 105 (48) (2008) 18970–18975.
[7] M. Hermans, B. Schrauwen, Memory in linear recurrent neural networks in continuous time, Neural Netw.: Off. J. Int. Neural Netw. Soc. 23 (3) (2010) 341–355.
[8] J. Dambre, D. Verstraeten, B. Schrauwen, S. Massar, Information processing capacity of dynamical systems, Sci. Rep. 2 (514) (2012).
[9] P. Barancok, 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, 2014, pp. 41–48.
[10] R. Couillet, G. Wainrib, H. Sevi, H.T. Ali, The asymptotic performance of linear echo state neural networks, J. Mach. Learn. Res. 17 (178) (2016) 1–35.
[11] I. Farkas, R. Bosak, P. Gergel, Computational analysis of memory capacity in echo state networks, Neural Netw. 83 (2016) 109–120.
[12] A. Goudarzi, S. Marzen, P. Banda, G. Feldman, M.R. Lakin, C. Teuscher, D. Stefanovic, Memory and Information Processing in Recurrent Neural Networks, Tech. Rep., 2016.
[13] F. Xue, Q. Li, 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) (2017) e0181816.
[14] P. Verzelli, C. Alippi, L. Livi, Echo state networks with self-normalizing activations on the hyper-sphere, Sci. Rep. 9 (1) (2019) 13887.
[15] L. Grigoryeva, J. Henriques, L. Larger, J.-P. Ortega, Optimal nonlinear information processing capacity in delay-based reservoir computers, Sci. Rep. 5 (12858) (2015) 1–11.
[16] L. Grigoryeva, J. Henriques, L. Larger, J.-P. Ortega, Nonlinear memory capacity of parallel time-delay reservoir computers in the processing of multidimensional signals, Neural Comput. 28 (2016) 1411–1451.
[17] S. Ortin, L. Pesquera, J.M. Gutiérrez, Memory and nonlinear mapping in reservoir computing with two uncoupled nonlinear delay nodes, in: T. Gilbert, M. Kirkilionis, G. Nicolis (Eds.), Proceedings of the European Conference on Complex Systems, Springer International Publishing Switzerland, 2012, pp. 895–899.
[18] L. Grigoryeva, J. Henriques, L. Larger, J.-P. Ortega, Stochastic time series forecasting using time-delay reservoir computers: performance and universality, Neural Netw. 55 (2014) 59–71.
[19] S. Ortín, L. Pesquera, Tackling the trade-off between information processing capacity and rate in delay-based reservoir computers, Front. Phys. 7 (2019) 210.
[20] S. Ortín, L. Pesquera, Delay-based reservoir computing: tackling performance degradation due to system response time, Opt. Lett. 45 (4) (2020) 905–908.
[21] P. Tino, A. Rodan, Short term memory in input-driven linear dynamical systems, Neurocomputing 112 (2013) 58–63.
[22] L. Livi, F.M. Bianchi, C. Alippi, Determination of the edge of criticality in echo state networks through Fisher information maximization, 2016.
[23] P. Tino, Asymptotic Fisher memory of randomized linear symmetric Echo State Networks, Neurocomputing 298 (2018) 4–8.
[24] A. Charles, H. Yap, C. Rozell, Short term network memory capacity via the restricted isometry property, Neural Comput. 26 (2014).
[25] L. Grigoryeva, J. Henriques, J.-P. Ortega, Reservoir computing: information processing of stationary signals, in: Proceedings of the 19th IEEE International Conference on Computational Science and Engineering, 2016, pp. 496–503.
[26] A.S. Charles, D. Yin, C.J. Rozell, Distributed Sequence Memory of Multidimensional Inputs in Recurrent Networks, Tech. Rep., 2017.
[27] S. Marzen, Difference between memory and prediction in linear recurrent networks, Phys. Rev. E 96 (3) (2017) 1–7.
[28] R. Kalman, Lectures on controllability and observability, in: Controllability and Observability, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 1–149.
[29] E.D. Sontag, Kalman's controllability rank condition: from linear to nonlinear, in: A.C. Antoulas (Ed.), Mathematical System Theory, Springer, 1991, pp. 453–462.
[30] E. Sontag, Mathematical Control Theory: Deterministic Finite Dimensional Systems, Springer-Verlag, 1998.
[31] A. Rodan, P. Tino, Minimum complexity echo state network, IEEE Trans. Neural Netw. 22 (1) (2011) 131–144.
[32] P.V. Aceituno, Y. Gang, Y.-Y. Liu, Tailoring artificial neural networks for optimal learning, 2017.
[33] P. Verzelli, C. Alippi, L. Livi, P. Tino, Input representation in recurrent neural networks dynamics, 2020.
[34] L. Grigoryeva, J.-P. Ortega, Echo state networks are universal, Neural Netw. 108 (2018) 495–508.
[35] M.B. Matthews, Approximating nonlinear fading-memory operators using neural network models, Circuits Systems Signal Process. 12 (2) (1993) 279–307.
[36] Lukas Gonon, Juan-Pablo Ortega, Reservoir computing universality with stochastic inputs, IEEE Transactions on Neural Networks and Learning Systems (2018).
[37] H. Jaeger, The 'Echo State' Approach to Analysing and Training Recurrent Neural Networks with an Erratum Note, Tech. Rep., German National Research Center for Information Technology, 2010.
[38] M. Buehner, P. Young, A tighter bound for the echo state property, IEEE Trans. Neural Netw. 17 (3) (2006) 820–824.
[39] I.B. Yildiz, H. Jaeger, S.J. Kiebel, Re-visiting the echo state property, Neural Netw. 35 (2012) 1–9.
[40] Bai Zhang, D.J. Miller, Yue Wang, Nonlinear system modeling with random matrices: echo state networks revisited, IEEE Trans. Neural Netw. Learn. Syst. 23 (1) (2012) 175–182.
[41] G. Wainrib, M.N. Galtier, A local echo state property through the largest Lyapunov exponent, Neural Netw. 76 (2016) 39–45.
[42] G. Manjunath, H. Jaeger, Echo state property linked to an input: exploring a fundamental characteristic of recurrent neural networks, Neural Comput. 25 (3) (2013) 671–696.
[43] C. Gallicchio, A. Micheli, Echo state property of deep reservoir computing networks, Cogn. Comput. 9 (2017).
[44] Lyudmila Grigoryeva, Juan-Pablo Ortega, Universal discrete-time reservoir computers with stochastic inputs and linear readouts using non-homogeneous state-affine systems, Journal of Machine Learning Research 19 (24) (2018) 1–40.
[45] L. Grigoryeva, J.-P. Ortega, Differentiable reservoir computing, J. Mach. Learn. Res. 20 (179) (2019) 1–62.
[46] L. Gonon, L. Grigoryeva, J.-P. Ortega, Risk bounds for reservoir computing, 2019, Preprint.
[47] J. Munkres, Topology, second ed., Pearson, 2014.
[48] P.J. Brockwell, R.A. Davis, Time Series: Theory and Methods, Springer-Verlag, 2006.
[49] P. Tino, Dynamical systems as temporal feature spaces, J. Mach. Learn. Res. 21 (2020) 1–42.
[50] O. Kallenberg, Foundations of Modern Probability, in: Probability and Its Applications, Springer New York, 2002.
[51] B.N. Mukherjee, S.S. Maiti, On some properties of positive definite Toeplitz matrices and their possible applications, Linear Algebra Appl. 102 (1988) 211–240.
[52] R.A. Horn, C.R. Johnson, Matrix Analysis, second ed., Cambridge University Press, 2013.
[53] R.M. Gray, Toeplitz and circulant matrices: A review, Found. Trends Commun. Inf. Theory 2 (3) (2006) 155–239.
[54] R.A. Horn, C.R. Johnson, Topics in Matrix Analysis, Cambridge University Press, Cambridge, 1994.