

Memory and forecasting capacities of nonlinear recurrent networks

Lukas Gonon¹, Lyudmila Grigoryeva², and Juan-Pablo Ortega^{3,4}

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 network and the autocovariance function of the input. These bounds generalize well-known estimates in the literature to a dependent inputs setup. Finally, for linear recurrent networks and independent inputs it is proved that the memory capacity is given by the rank of the associated controllability matrix.

Key Words: memory capacity, forecasting capacity, recurrent neural network, reservoir computing, echo state network, ESN, linear recurrent network, machine learning, fading memory property, echo state property.

1 Introduction

Memory capacities have been introduced in [Jaeg 02] in the context of recurrent neural networks in general and of echo state networks (ESNs) [Matt 92, Matt 94, Jaeg 04] 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.

In the original definition, the memory capacity was defined as the sum of the coefficients of determination of the different linear regressions that take the state of the system at a given time as covariates and values of the input at a given point in the past as dependent variables. This notion has been the subject of much research in the reservoir computing literature [Whit 04, Gang 08, Herm 10, Damb 12, Bara 14, Coui 16, Fark 16, Goud 16, Char 17, Xue 17, Verz 19] where most of the efforts have been concentrated in linear and echo state systems. Analytical expression of the capacity of time-delay reservoirs have been formulated in [Grig 15, Grig 16a] and various proposals for optimized reservoir architectures can be obtained by maximizing this parameter [Orti 12, Grig 14, Orti 19, Orti 20]. Additionally, memory capacities have been extensively compared with other related concepts like Fisher information-base criteria [Tino 13, Livi 16, Tino 18].

¹Universität Sankt Gallen. Faculty of Mathematics and Statistics. Bodanstrasse 6. CH-9000 Sankt Gallen. Switzerland. lukas.gonon@unisg.ch

²Department of Mathematics and Statistics. Universität Konstanz. Box 146. D-78457 Konstanz. Germany. Lyudmila.Grigoryeva@uni-konstanz.de

³Universität Sankt Gallen. Faculty of Mathematics and Statistics. Bodanstrasse 6. CH-9000 Sankt Gallen. Switzerland. Juan-Pablo.Ortega@unisg.ch

⁴Centre National de la Recherche Scientifique (CNRS). France.

All the above-mentioned works consider exclusively white noise as input signals and it is, to our knowledge only in [Grig 16b] (under strong high-order stationarity assumptions) and in [Marz 17] (for linear recurrent networks) 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 [Marz 17] and studied in detail for linear recurrent networks. This work shows, in particular, that linear networks optimized for memory capacity do not necessarily have a good forecasting capacity and vice versa.

The main contribution of this paper is the extension of the results on memory and forecasting capacities available in the literature for either linear recurrent or echo state networks, or for independent inputs, to general non-linear systems and to dependent inputs that are only assumed to be stationary. More explicitly, it is known since [Jaeg 02] that the memory capacity of an ESN or a linear recurrent network defined using independent inputs is bounded above by its number of output neurons or, equivalently, by the dimensionality of the corresponding state-space representation. Moreover, in the linear case, it can be shown [Jaeg 02] that this capacity is maximal if and only if Kalman’s controllability rank condition [Kalm 10, Sont 91, Sont 98] is satisfied. This paper extends those results as follows:

- Theorem 3.4 states generic bounds for the memory and the forecasting capacities of generic recurrent network with linear readouts and with stationary inputs, in terms of the dimensionality of the corresponding state-space representation and the autocovariance function of the input, which is assumed to be second-order stationary. These bounds reduce to those in [Jaeg 02] when the inputs are independent.
- Section 4 is exclusively devoted to the linear case. In that simpler situation, explicit expressions for the memory and forecasting capacities can be stated (see Proposition 4.1) in terms of the matrix parameters of the network and the autocorrelation properties of the input.
- In the *classical* linear case with independent inputs, the expressions in the previous point 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 system [Kalm 10]. This last condition has been already mentioned in [Jaeg 02] in relation with maximum capacity.
- 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*. The relation between maximum capacity of a system and the full rank of the controllability matrix established in [Jaeg 02] is obviously a particular case of this statement. 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 ingenious configurations that maximize that rank [Acei 17, Verz 20].

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:

$$\begin{cases} \mathbf{x}_t = F(\mathbf{x}_{t-1}, \mathbf{z}_t), & (2.1) \\ \mathbf{y}_t = h(\mathbf{x}_t) := \mathbf{W}^\top \mathbf{x}_t + \mathbf{a}. & (2.2) \end{cases}$$

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 $\mathbf{z}_t \in \mathbb{R}^d$, will be in most cases infinite paths of a discrete-time stochastic process.

We shall focus on state-space systems of the type (2.1)-(2.2) that determine an **input/output** system. This happens in the presence of the so-called **echo state property (ESP)**, that is, when for any $\mathbf{z} \in (D_d)^{\mathbb{Z}}$ there exists a unique $\mathbf{y} \in (\mathbb{R}^m)^{\mathbb{Z}}$ such that (2.1)-(2.2) hold. In that case, we talk about the **state-space filter** $U_h^F : (D_d)^{\mathbb{Z}} \rightarrow (\mathbb{R}^m)^{\mathbb{Z}}$ associated to the state-space system (2.1)-(2.2) defined by:

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

with $\mathbf{z} \in (D_d)^{\mathbb{Z}}$ and $\mathbf{y} \in (\mathbb{R}^m)^{\mathbb{Z}}$ are linked by (2.1) via the ESP. If the ESP holds at the level of the state equation (2.1), we can define a **state filter** $U^F : (D_d)^{\mathbb{Z}} \rightarrow (D_N)^{\mathbb{Z}}$ 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 [Grig 18, Proposition 2.1]) and hence it suffices to work with their restriction $U_h^F : (D_d)^{\mathbb{Z}_-} \rightarrow (\mathbb{R}^m)^{\mathbb{Z}_-}$ to semi-infinite inputs and outputs. Moreover, U_h^F determines a state-space **functional** $H_h^F : (D_d)^{\mathbb{Z}} \rightarrow \mathbb{R}^m$ 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 deserved much attention in the context of the so-called **echo state networks (ESNs)** [Matt 92, Matt 93, Matt 94, Jaeg 04] (see, for instance, [Jaeg 10, Bueh 06, Yild 12, Bai 12, Wain 16, Manj 13, Gall 17]). Sufficient conditions for the ESP to hold in general systems have been formulated in [Grig 18, Grig 19, Gono 19], in most cases assuming that the state map is a contraction in the state variable. We recall now a result (see [Grig 19, Theorem 12] and [Gono 19, Proposition 1]) that ensures the ESP as well as a continuity property of importance for the sequel.

Proposition 2.1 *Let $F : D_N \times D_d \rightarrow D_N$ 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 the echo state property for any input in $(D_n)^{\mathbb{Z}_-}$. The associated filter $U^F : (D_n)^{\mathbb{Z}_-} \rightarrow (D_N)^{\mathbb{Z}_-}$ is continuous with respect to the product topologies in $(D_n)^{\mathbb{Z}_-}$ and $(D_N)^{\mathbb{Z}_-}$.

State-space morphisms. In this paragraph we discuss how maps between state spaces can produce different state-space representations for a given filter. Consider the state-space systems determined by the two pairs (F_i, h_i) , $i \in \{1, 2\}$, with $F_i : D_{N_i} \times D_d \rightarrow D_{N_i}$ and $h_i : D_{N_i} \rightarrow \mathbb{R}^m$.

Definition 2.2 *A map $f : D_{N_1} \rightarrow D_{N_2}$ 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 it is also a morphism between the systems determined by the pairs (F_1, M_1) and (F_2, M_2) we say that f is a **system isomorphism** and that the systems (F_1, M_1) and (F_2, M_2) are **isomorphic**. Given a system $F_1 : D_{N_1} \times D_d \rightarrow D_{N_1}$, $h_1 : D_{N_1} \rightarrow$

\mathbb{R}^m and a bijection $f : D_{N_1} \rightarrow D_{N_2}$, the map f is a system isomorphism with respect to the system $F_2 : D_{N_2} \times D_d \rightarrow D_{N_2}, h_2 : D_{N_2} \rightarrow \mathbb{R}^m$ defined by

$$F_2(\mathbf{x}_2, \mathbf{z}) := f(F_1(f^{-1}(\mathbf{x}_2), \mathbf{z})), \quad \text{for all } \mathbf{x}_2 \in D_{N_2}, \mathbf{z} \in D_d, \quad (2.3)$$

$$h_2(\mathbf{x}_2) := h_1(f^{-1}(\mathbf{x}_2)), \quad \text{for all } \mathbf{x}_2 \in D_{N_2}. \quad (2.4)$$

Proposition 2.3 *Let (F_1, h_1) and (F_2, h_2) be the two systems introduced above. Let $f : D_{N_1} \rightarrow D_{N_2}$ 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 $(f(\mathbf{x}_t^1))_{t \in \mathbb{Z}_-} \in (D_{N_2})^{\mathbb{Z}_-}$ 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,*

$$U_{h_1}^{F_1} = U_{h_2}^{F_2}. \quad (2.5)$$

- (iii) *If f is a system isomorphism then the implications in the previous two points are reversible.*

Proof. (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 (2.5) holds, it suffices to prove that given any $\mathbf{z} \in (D_d)^{\mathbb{Z}_-}$, any solution $(\mathbf{y}^1, \mathbf{x}^1) \in (\mathbb{R}^m)^{\mathbb{Z}_-} \times (D_{N_1})^{\mathbb{Z}_-}$ of (F_1, h_1) associated to \mathbf{z} , with $\mathbf{y}^1 := (h_1(\mathbf{x}_t^1))_{t \in \mathbb{Z}_-}$, which exists by the hypothesis on F_1 , coincides with the unique solution $U_{h_2}^{F_2}(\mathbf{z})$ for the system (F_2, h_2) . Indeed, for any $t \in \mathbb{Z}_-$,

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

Here, the first equality follows from readout invariance and the second one from the system equivariance. This implies that $(\mathbf{y}^1, (f(\mathbf{x}_t^1))_{t \in \mathbb{Z}_-})$ 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 $\mathbf{y}^1 = U_{h_2}^{F_2}(\mathbf{z})$ and since $\mathbf{z} \in V_d$ is arbitrary, the result follows. Part (iii) is straightforward. ■

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 $\mathbb{P} : \mathcal{A} \rightarrow [0, 1]$. The input and target signals are modeled by discrete-time stochastic processes $\mathbf{Z} = (\mathbf{Z}_t)_{t \in \mathbb{Z}_-}$ and $\mathbf{Y} = (\mathbf{Y}_t)_{t \in \mathbb{Z}_-}$ taking values in $D_d \subset \mathbb{R}^d$ and \mathbb{R}^m , respectively. Moreover, we write $\mathbf{Z}(\omega) = (\mathbf{Z}_t(\omega))_{t \in \mathbb{Z}_-}$ and $\mathbf{Y}(\omega) = (\mathbf{Y}_t(\omega))_{t \in \mathbb{Z}_-}$ for each outcome $\omega \in \Omega$ to denote the realizations or sample paths of \mathbf{Z} and \mathbf{Y} , respectively. Since \mathbf{Z} can be seen as a random sequence in $D_d \subset \mathbb{R}^d$, we write interchangeably $\mathbf{Z} : \mathbb{Z}_- \times \Omega \rightarrow D_d$ and $\mathbf{Z} : \Omega \rightarrow (D_d)^{\mathbb{Z}_-}$. The latter is necessarily measurable with respect to the Borel σ -algebra induced by the product topology in $(D_d)^{\mathbb{Z}_-}$. The same applies to the analogous assignments involving \mathbf{Y} .

We will most of the time work under *stationarity* hypotheses. We recall that the discrete-time process $\mathbf{Z} : \Omega \rightarrow (D_d)^{\mathbb{Z}_-}$ is stationary whenever $T_{-\tau}(\mathbf{Z}) \stackrel{d}{=} \mathbf{Z}$, for any $\tau \in \mathbb{Z}_-$, where the symbol $\stackrel{d}{=}$

stands for the equality in distribution and $T_{-\tau} : (\mathbb{R}^d)^{\mathbb{Z}_-} \rightarrow (\mathbb{R}^d)^{\mathbb{Z}_-}$ 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** [Broc 06]. 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-summable process $\mathbf{Z} : \Omega \rightarrow (D_d)^{\mathbb{Z}_-}$ is second-order stationary whenever (i) there exists a constant $\boldsymbol{\mu} \in \mathbb{R}^d$ such that $\mathbb{E}[\mathbf{Z}_t] = \boldsymbol{\mu}$, for all $t \in \mathbb{Z}_-$ (**mean stationarity**) and (ii) the autocovariance matrices $\text{Cov}(\mathbf{Z}_t, \mathbf{Z}_{t+h})$ depend only on $h \in \mathbb{Z}$ and not on $t \in \mathbb{Z}_-$ (**autocovariance stationarity**) and we can hence define the **autocovariance function** $\gamma : \mathbb{Z} \rightarrow \mathbb{R}$ as $\gamma(h) := \text{Cov}(\mathbf{Z}_t, \mathbf{Z}_{t+h})$, which is necessarily even [Broc 06, Proposition 1.5.1]. 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 $F : D_N \times D_d \rightarrow D_N$ be a state map that satisfies the hypotheses of Proposition 2.1 or that, more generally, it has the echo state property and the associated filter $U^F : (D_n)^{\mathbb{Z}_-} \rightarrow (D_N)^{\mathbb{Z}_-}$ is continuous with respect to the product topologies in $(D_n)^{\mathbb{Z}_-}$ and $(D_N)^{\mathbb{Z}_-}$. If the input process $\mathbf{Z} : \Omega \rightarrow (D_d)^{\mathbb{Z}_-}$ is stationary, then so is the output $\mathbf{X} := U^F(\mathbf{Z}) : \Omega \rightarrow (D_N)^{\mathbb{Z}_-}$ as well as the joint processes $(T_{-\tau}(\mathbf{X}), \mathbf{Z})$ and $(\mathbf{X}, T_{-\tau}(\mathbf{Z}))$, for any $\tau \in \mathbb{Z}_-$.*

Proof. The continuity (and hence the measurability) hypothesis on U^F proves that $\mathbf{X} := U^F(\mathbf{Z})$ is stationary (see [Kall 02, 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. ■

The next result shows that if a state-space system has covariance stationary outputs and the corresponding covariance matrix is non-singular, then there exists an isomorphic state-space representation whose states have mean zero and have orthonormal components. The stationarity condition in this result can be obtained, for example, by using Corollary 2.4.

Proposition 2.5 (Standardization of state-space realizations) *Consider a state-space system as in (2.1)-(2.2) and suppose that the input process $\mathbf{Z} : \Omega \rightarrow (D_d)^{\mathbb{Z}_-}$ is such that the associated state process $\mathbf{X} : \Omega \rightarrow (D_N)^{\mathbb{Z}_-}$ is covariance stationary. Let $\boldsymbol{\mu} := \mathbb{E}[\mathbf{X}_t]$ and suppose that the covariance matrix $\Gamma_{\mathbf{X}} := \text{Cov}(\mathbf{X}_t, \mathbf{X}_t)$ is non-singular. Then, the map $f : \mathbb{R}^N \rightarrow \mathbb{R}^N$ given by $f(\mathbf{x}) := \Gamma_{\mathbf{X}}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ is a system isomorphism between the system (2.1)-(2.2) and the one with state map*

$$\tilde{F}(\mathbf{x}, z) := \Gamma_{\mathbf{X}}^{-1/2} \left(F \left(\Gamma_{\mathbf{X}}^{1/2} \mathbf{x} + \boldsymbol{\mu}, z \right) - \boldsymbol{\mu} \right) \quad (2.6)$$

and readout

$$\tilde{h}(\mathbf{x}) := (\Gamma_{\mathbf{X}}^{-1/2} \mathbf{W})^\top \mathbf{x} + \mathbf{W}^\top \boldsymbol{\mu} + \mathbf{a}. \quad (2.7)$$

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

$$\mathbb{E}[\tilde{\mathbf{X}}_t] = \mathbf{0}, \quad \text{and} \quad \text{Cov}(\tilde{\mathbf{X}}_t, \tilde{\mathbf{X}}_t) = \mathbb{I}_N. \quad (2.8)$$

Proof. 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 (\tilde{F}, \tilde{h}) is a consequence of the equalities (2.3)-(2.4). 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 $\tilde{\mathbf{X}}$ defined by $\tilde{\mathbf{X}}_t := \Gamma_{\mathbf{X}}^{-1/2}(\mathbf{X}_t - \boldsymbol{\mu})$, $t \in \mathbb{Z}_-$, with respect to \tilde{F} . The equalities (2.8) immediately follow. ■

3 Memory and forecasting capacity bounds for stationary inputs

The following definition extends the notion of memory capacity introduced in [Jaeg 02] to general nonlinear systems and to input signals that are stationary but not necessarily time-decorrelated.

Definition 3.1 Let $\mathbf{Z} : \Omega \rightarrow D^{\mathbb{Z}^-}$, $D \subset \mathbb{R}$, 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 $\mathbf{X} : \Omega \rightarrow (D_N)^{\mathbb{Z}^-}$ 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:

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

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

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

$$\text{MC} := \sum_{\tau \in \mathbb{Z}^-} \text{MC}_\tau, \quad \text{FC} := \sum_{\tau \in \mathbb{Z}^-} \text{FC}_\tau. \quad (3.3)$$

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 proof of the following result is straightforward.

Lemma 3.2 In the conditions of Definition (3.1) and if the covariance matrix $\Gamma_{\mathbf{X}} := \text{Cov}(\mathbf{X}_t, \mathbf{X}_t)$ is non-singular then, for any $\tau \in \mathbb{Z}^-$:

$$\text{MC}_\tau = \frac{\text{Cov}(Z_{t+\tau}, \mathbf{X}_t) \Gamma_{\mathbf{X}}^{-1} \text{Cov}(\mathbf{X}_t, Z_{t+\tau})}{\text{Var}(Z_t)}, \quad \text{FC}_\tau = \frac{\text{Cov}(Z_t, \mathbf{X}_{t+\tau}) \Gamma_{\mathbf{X}}^{-1} \text{Cov}(\mathbf{X}_{t+\tau}, Z_t)}{\text{Var}(Z_t)}. \quad (3.4)$$

The next result shows that the memory capacity of a state-space system with linear readouts is invariant with respect to linear system morphisms. This result will be used later on in Section 4 when we study the memory and forecasting capacities of linear systems. The proof uses 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 $f : V \rightarrow W$ be a linear map between them. The dual map $f^* : W \rightarrow V$ of f is defined by

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

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

$$\langle \mathbf{v}_1, \mathbf{v}_2 \rangle_V := \langle f(\mathbf{v}_1), f(\mathbf{v}_2) \rangle_W. \quad (3.5)$$

It is easy to see that in that case, the dual map with respect to the inner product (3.5) in V and $\langle \cdot, \cdot \rangle_W$ in W , satisfies that

$$f^* \circ f = \mathbb{I}_W, \quad (3.6)$$

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

Lemma 3.3 *Let \mathbf{Z} be a variance-stationary input and let $F_2 : D_{N_2} \times D \rightarrow D_{N_2}$ be a state map that satisfies the conditions of Definition 3.1. Let $F_1 : D_{N_1} \times D \rightarrow D_{N_1}$ be another state map and let $f : \mathbb{R}^{N_1} \rightarrow \mathbb{R}^{N_2}$ 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} .*

Proof. 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 $U_{\mathbf{W} \circ f}^{F_1}$ and

$$U_{\mathbf{W} \circ f}^{F_1} = U_{\mathbf{W}}^{F_2}. \quad (3.7)$$

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 (3.5), and let f^* be the corresponding dual map. It is easy to see that the equality (3.7) can be rewritten using f^* as

$$U_{f^*(\mathbf{W})}^{F_1} = U_{\mathbf{W}}^{F_2}, \quad \text{for any } \mathbf{W} \in \mathbb{R}^{N_2}. \quad (3.8)$$

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 (3.8)

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

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 (3.6)). A similar statement can be written for the forecasting capacities. ■

The next result is the main contribution of this paper and generalizes the bounds formulated in [Jaeg 02] 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.

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

(i) *If the input process \mathbf{Z} is mean-stationary then, for any $\tau \in \mathbb{Z}_-$:*

$$0 \leq \text{MC}_\tau \leq 1, \quad \text{and} \quad 0 \leq \text{FC}_\tau \leq 1. \quad (3.9)$$

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

$$0 \leq C \leq \frac{N}{\gamma(0)} \rho(H) \leq N \left(1 + \frac{2}{\gamma(0)} \sum_{j=1}^{\infty} |\gamma(j)| \right), \quad (3.10)$$

where $\rho(H) := \lim_{L \rightarrow -\infty} \rho(H^L)$, 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 $f : [-\pi, \pi] \rightarrow \mathbb{R}$ of \mathbf{Z} is well-defined and given by

$$f(\lambda) := \frac{1}{2\pi} \sum_{n \in \mathbb{Z}} e^{-in\lambda} \gamma(n), \quad (3.11)$$

and we have that

$$0 \leq C \leq \frac{2\pi N}{\gamma(0)} M_f \leq N \left(1 + \frac{2}{\gamma(0)} \sum_{j=1}^{\infty} |\gamma(j)| \right), \quad (3.12)$$

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

Proof. 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 (2.1)-(2.2) in the definitions (3.1) and (3.2) by its standardized counterpart whose states $\tilde{\mathbf{X}}_t$ are such that $\mathbb{E}[\tilde{\mathbf{X}}_t] = \mathbf{0}$ and $\text{Cov}(\tilde{\mathbf{X}}_t, \tilde{\mathbf{X}}_t) = \mathbb{E}[\tilde{\mathbf{X}}_t \tilde{\mathbf{X}}_t^\top] = \mathbb{I}_N$. More specifically, if we denote by $\sigma^2 := \text{Var}(Z_t) = \gamma(0)$, we can write

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

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

$$\text{MC}_\tau = \frac{1}{\sigma^2} \text{Cov}(Z_{t+\tau}, \tilde{\mathbf{X}}_t) \text{Cov}(\tilde{\mathbf{X}}_t, Z_{t+\tau}) = \frac{1}{\sigma^2} \sum_{i=1}^N \mathbb{E} \left[\tilde{X}_t^i Z_{t+\tau} \right]^2. \quad (3.13)$$

Analogously, it is easy to show that

$$\text{FC}_\tau = \frac{1}{\sigma^2} \sum_{i=1}^N \mathbb{E} \left[\tilde{X}_{t+\tau}^i Z_t \right]^2. \quad (3.14)$$

If we now define $\tilde{Z}_t := Z_t/\sigma \in L^2(\Omega, \mathbb{R})$, for any $t \in \mathbb{Z}_-$, it is clear that $\|\tilde{Z}_t\|_{L^2} = \mathbb{E}[\tilde{Z}_t^2] = 1$. Moreover, the relation $\text{Cov}(\tilde{\mathbf{X}}_t, \tilde{\mathbf{X}}_t) = \mathbb{I}_N$ implies that the components $\tilde{X}_t^i \in L^2(\Omega, \mathbb{R})$, $i \in \{1, \dots, N\}$, and that they form an orthonormal set, that is, $\langle \tilde{X}_t^i, \tilde{X}_t^j \rangle_{L^2} = \delta_{ij}$, where δ_{ij} stands for Kronecker's delta. We now separately prove the two parts of the theorem.

(i) The fact that $\text{MC}_\tau, \text{FC}_\tau \geq 0$ is obvious from (3.13) and (3.14). Let now $\tilde{S}_t = \text{span} \{ \tilde{X}_t^1, \dots, \tilde{X}_t^N \} \subset L^2(\Omega, \mathbb{R})$ and let $\mathbb{P}_{\tilde{S}_t} : L^2(\Omega, \mathbb{R}) \rightarrow \tilde{S}_t$ be the corresponding orthogonal projection. Then,

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

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

(ii) Define first, for any $L \in \mathbb{Z}_-$, the vector $\mathbf{Z}^L := (Z_0, Z_{-1}, \dots, Z_L)$, and

$$\text{MC}^L := \sum_{\tau=0}^L \text{MC}_\tau = \sum_{\tau=0}^L \sum_{i=1}^N \mathbb{E} \left[\tilde{X}_0^i \tilde{Z}_\tau \right]^2, \quad \text{FC}^L := \sum_{\tau=-1}^L \text{FC}_\tau = \sum_{\tau=-1}^L \sum_{i=1}^N \mathbb{E} \left[\tilde{X}_L^i \tilde{Z}_{L-\tau} \right]^2. \quad (3.15)$$

In these equalities we used (3.13) and (3.14) as well as the stationarity hypothesis. Now, using the invertibility hypothesis on the matrices H^L , we define the random vector $\widehat{\mathbf{Z}}^L := (H^L)^{-1/2} \mathbf{Z}^L$, whose components form an orthonormal set in $L^2(\Omega, \mathbb{R})$. Indeed, for any $i, j \in \{1, \dots, -L+1\}$,

$$\begin{aligned} \langle \widehat{Z}_i^L, \widehat{Z}_j^L \rangle_{L^2} &= \mathbb{E} \left[\widehat{Z}_i^L \widehat{Z}_j^L \right] = \sum_{k,l=1}^{-L+1} (H^L)^{-1/2}_{ik} (H^L)^{-1/2}_{jl} \mathbb{E} [Z_{-k+1} Z_{-l+1}] \\ &= \sum_{k,l=1}^{-L+1} (H^L)^{-1/2}_{ik} (H^L)^{-1/2}_{jl} \gamma(|k-l|) = \left((H^L)^{-1/2} H^L (H^L)^{-1/2} \right)_{ij} = \delta_{ij}. \end{aligned} \quad (3.16)$$

Let $\widetilde{S}_L = \text{span} \left\{ \widehat{Z}_1^L, \dots, \widehat{Z}_{-L+1}^L \right\} \subset L^2(\Omega, \mathbb{R})$ and let $\mathbb{P}_{\widetilde{S}_L} : L^2(\Omega, \mathbb{R}) \rightarrow \widetilde{S}_L$ be the corresponding orthogonal projection. By (3.15) we have that

$$\text{MC}^L = \sum_{\tau=0}^L \sum_{i=1}^N \mathbb{E} \left[\widetilde{X}_0^i \widetilde{Z}_\tau \right]^2 = \frac{1}{\gamma(0)} \sum_{\tau=0}^L \sum_{i=1}^N \mathbb{E} \left[\widetilde{X}_0^i \left((H^L)^{1/2} \widehat{\mathbf{Z}}^L \right)_{-\tau+1} \right]^2 = \frac{1}{\gamma(0)} \sum_{i=1}^N \left\| (H^L)^{1/2} \mathbb{E} \left[\widetilde{X}_0^i \widehat{\mathbf{Z}}^L \right] \right\|^2. \quad (3.17)$$

Analogously,

$$\begin{aligned} \text{FC}^L &= \sum_{\tau=-1}^L \sum_{i=1}^N \mathbb{E} \left[\widetilde{X}_L^i \widetilde{Z}_{L-\tau} \right]^2 = \frac{1}{\gamma(0)} \sum_{\tau=-1}^L \sum_{i=1}^N \mathbb{E} \left[\widetilde{X}_L^i \left((H^L)^{1/2} \widehat{\mathbf{Z}}^L \right)_{-L+\tau+1} \right]^2 \\ &\leq \frac{1}{\gamma(0)} \sum_{i=1}^N \left\| (H^L)^{1/2} \mathbb{E} \left[\widetilde{X}_L^i \widehat{\mathbf{Z}}^L \right] \right\|^2. \end{aligned} \quad (3.18)$$

Now, by (3.16) and (3.17) we can write that

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

An identical inequality can be shown for FC^L using (3.18), which proves the first inequality in (3.10). The second inequality can be obtained by bounding $\rho(H^L)$ using Gersgorin's Disks Theorem (see [Horn 13, Theorem 6.1.1 and Corollary 6.1.5]). Indeed, due to this result:

$$\rho(H^L) \leq \gamma(0) + \max_{i \in \{1, \dots, -L+1\}} \left\{ \sum_{\substack{j \in \{1, \dots, -L+1\} \\ j \neq i}} |H_{ij}^L| \right\} \leq \gamma(0) + 2 \sum_{i=1}^{-L} |\gamma(i)|. \quad (3.19)$$

(iii) The inequalities in (3.12) 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 (3.11) determines the spectral density of \mathbf{Z} by [Broc 06, Corollary 4.3.2]. Second, by [Gray 06, Lemma 6, page 194], the spectrum of H is bounded above by the

maximum of the function $2\pi f$, which, using (3.10) implies that $C \leq \frac{2\pi N}{\gamma(0)} M_f$. The last inequality is a consequence of

$$|f(\lambda)| \leq \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} |\gamma(j)| = \left(\gamma(0) + 2 \sum_{j=1}^{\infty} |\gamma(j)| \right) < +\infty, \quad \text{for any } \lambda \in [-\pi, \pi]. \quad \blacksquare$$

This result and its proof are used in the next corollary to recover the memory capacity bounds proposed in [Jaeg 02] when using independent inputs and to show that, in that case, the forecasting capacity is zero.

Corollary 3.5 (Jaeger [Jaeg 02]) *In the conditions of Theorem 3.4, if the input process \mathbf{Z} is independent, then*

$$0 \leq \text{MC} \leq N \quad \text{and} \quad \text{FC} = 0. \quad (3.20)$$

Proof. The first inequality is a straightforward consequence of (3.10) and the fact that for independent inputs $\gamma(h) = 0$, for all $h \neq 0$. The second one can be easily obtained from (3.14). Indeed, by the causality and the time-invariance [Grig 18, Proposition 2.1] of any filter induced by a state-space system of the type (2.1)-(2.2), for any $\tau \leq -1$, the random variables $\tilde{X}_{t+\tau}^i$ and Z_t in (3.14) are independent, and hence

$$\text{FC}_\tau = \frac{1}{\sigma^2} \sum_{i=1}^N \text{E} \left[\tilde{X}_{t+\tau}^i Z_t \right]^2 = \frac{1}{\sigma^2} \sum_{i=1}^N \text{E} \left[\tilde{X}_{t+\tau}^i \right]^2 \text{E} [Z_t]^2 = 0. \quad \blacksquare$$

4 The memory and forecasting capacities of linear systems

When the state equation (2.1) 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 [Jaeg 02, Coui 16] and, more recently, in [Marz 17] for more general correlated inputs and diagonalizable linear systems. It is in this paper that has been pointed out for the first time how different linear systems that maximize forecasting and memory capacities may be.

We split this section 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 outputs. In the second one we see how, using Lemma 3.3, we can reduce the more 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*. This is a generalization of the fact, already established in [Jaeg 02], that when the rank of the controllability is maximal, then linear systems has maximal capacity, that is, it coincides with the dimensionality of its state space. Different configurations that maximize the rank of the controllability matrix have been recently studied in [Acei 17, Verz 20].

4.1 The regular case

The 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 outputs. In this case we consider fully infinite inputs, that is, we shall be working with stationary input process $\mathbf{Z} : \Omega \rightarrow \mathbb{R}^{\mathbb{Z}}$ that have second-order moments and an autocovariance function $\gamma : \mathbb{Z} \rightarrow \mathbb{R}$ for which we shall construct the semi-infinite (respectively, doubly infinite) symmetric Toeplitz matrix $H_{ij} := \gamma(|i - j|)$, $i, j \in \mathbb{N}^+$, (respectively, $\bar{H}_{ij} := \gamma(|i - j|)$, $i, j \in \mathbb{Z}$).

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

$$F(\mathbf{x}, z) := \mathbf{A}\mathbf{x} + \mathbf{C}z, \quad \text{where } \mathbf{C} \in \mathbb{R}^N, \mathbf{A} \in \mathbb{M}_N, \text{ and } \|\mathbf{A}\|_2 = \sigma_{\max}(\mathbf{A}) < 1. \quad (4.1)$$

Consider a zero-mean stationary input process $\mathbf{Z} : \Omega \rightarrow D^{\mathbb{Z}}$, with $D \subset \mathbb{R}$ compact that has second-order moments and an absolutely summable autocovariance function $\gamma : \mathbb{Z} \rightarrow \mathbb{R}$. 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 $\mathbf{X} := U^{A, \mathbf{C}}(\mathbf{Z}) : \Omega \rightarrow (D_N)^{\mathbb{Z}_-}$ 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. Suppose that the covariance matrix $\Gamma_{\mathbf{X}} := \text{Cov}(\mathbf{X}_t, \mathbf{X}_t)$ is non-singular. Then,

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

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

These vectors form an orthonormal set in $\ell_+^2(\mathbb{R})$ and the memory capacity MC can be written as

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

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

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

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

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

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

Proof. First of all, the condition $\sigma_{\max}(\mathbf{A}) < 1$ in (4.1) implies that the state map is a contraction on the first entry. Second, as the input process takes values on a compact set, there exists a compact subset $D_N \subset \mathbb{R}^N$ (see [Gono 19, Remark 2]) such that the restriction $F : D_N \times D \rightarrow D_N$ 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,

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

and hence, using the notation introduced in Proposition 2.5 and the invertibility hypothesis on $\Gamma_{\mathbf{X}}$, the standardized states $\tilde{\mathbf{X}}_t$ are given by

$$\tilde{\mathbf{X}}_t = \Gamma_{\mathbf{X}}^{-1/2} \mathbf{X}_t = \Gamma_{\mathbf{X}}^{-1/2} \sum_{j=0}^{\infty} A^j \mathbf{C} \mathbf{Z}_{t-j}. \quad (4.6)$$

Additionally, when the autocovariance function is absolutely summable, then the spectral density f of \mathbf{Z} defined in (3.11) 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 \bar{H} (doubly infinite) are invertible (see [Gray 06, Theorem 11]).

(i) Let $\bar{\mathbf{Z}} := (Z_0, Z_1, \dots)$ and let $\hat{\mathbf{Z}} = H^{-1/2}\bar{\mathbf{Z}}$. An argument similar to (3.16) shows that $\langle \hat{Z}_i, \hat{Z}_j \rangle_{L^2} = \delta_{ij}$. Moreover, (4.6) implies that

$$\tilde{\mathbf{X}}_0 = \Gamma_{\mathbf{X}}^{-1/2} \sum_{k=0}^{\infty} A^k \mathbf{C} \mathbf{Z}_{-k} = \Gamma_{\mathbf{X}}^{-1/2} \sum_{k=0}^{\infty} A^k \mathbf{C} \left(H^{1/2} \hat{\mathbf{Z}} \right)_{k+1} = \Gamma_{\mathbf{X}}^{-1/2} \sum_{k=0}^{\infty} \sum_{j=1}^{\infty} A^k \mathbf{C} H_{k+1,j}^{1/2} \hat{Z}_j,$$

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

$$\tilde{X}_0^i = \sum_{j=1}^{\infty} B_i^j \hat{Z}_j, \quad i \in \{1, \dots, N\}.$$

The L^2 -orthonormality of the components \tilde{X}_0^i of $\tilde{\mathbf{X}}_0$ implies the ℓ^2 -orthonormality of the vectors $\{\mathbf{B}_1, \dots, \mathbf{B}_N\} \in \ell_+^2(\mathbb{R})$. Indeed, for any $i, j \in \{1, \dots, N\}$,

$$\delta_{ij} = \langle \tilde{X}_0^i, \tilde{X}_0^j \rangle_{L^2} = \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \mathbb{E} \left[\langle B_i^k \hat{Z}_k, B_j^l \hat{Z}_l \rangle_{L^2} \right] = \sum_{k=1}^{\infty} B_i^k B_j^k = \langle \mathbf{B}_i, \mathbf{B}_j \rangle_{\ell_+^2}. \quad (4.7)$$

We now prove (4.3). First, taking the limit $L \rightarrow \infty$ in (3.17) we write,

$$\begin{aligned} \text{MC} &= \frac{1}{\gamma(0)} \sum_{i=1}^N \left\| H^{1/2} \mathbb{E} \left[\tilde{X}_0^i \hat{\mathbf{Z}} \right] \right\|^2 = \frac{1}{\gamma(0)} \sum_{i=1}^N \left\| H^{1/2} \sum_{j=1}^{\infty} B_i^j \mathbb{E} \left[\hat{Z}_j \hat{\mathbf{Z}} \right] \right\|^2 \\ &= \frac{1}{\gamma(0)} \sum_{i=1}^N \left\| H^{1/2} \mathbf{B}_i \right\|_{\ell^2}^2 = \frac{1}{\gamma(0)} \sum_{i=1}^N \langle \mathbf{B}_i, H \mathbf{B}_i \rangle_{\ell^2}. \end{aligned}$$

(ii) Let now $\bar{\mathbf{Z}} := (\dots, Z_{-1}, Z_0, Z_1, \dots)$ and let $\hat{\mathbf{Z}} = \bar{H}^{-1/2}\bar{\mathbf{Z}}$. An argument similar to (3.16) shows that $\langle \hat{Z}_i, \hat{Z}_j \rangle_{L^2} = \delta_{ij}$ for any $i, j \in \mathbb{Z}$. Also, in this case, (4.6) implies that

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

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

$$\tilde{X}_0^i = \sum_{j=-\infty}^{\infty} B_i^j \hat{Z}_j, \quad i \in \{1, \dots, N\}.$$

As in (4.7), the L^2 -orthonormality of the components \tilde{X}_0^i of $\tilde{\mathbf{X}}_0$ implies the ℓ^2 -orthonormality of the vectors $\{\mathbf{B}_1, \dots, \mathbf{B}_N\} \in \ell^2(\mathbb{R})$. Finally, in order to establish (4.5), we use the stationarity hypothesis to

rewrite the expression of the forecasting capacity in (3.14) as

$$\begin{aligned}
\text{FC} &= \frac{1}{\gamma(0)} \sum_{\tau=1}^{\infty} \sum_{i=1}^N \mathbb{E} \left[\tilde{X}_0^i Z_{\tau} \right]^2 = \frac{1}{\gamma(0)} \sum_{\tau=1}^{\infty} \sum_{i=1}^N \mathbb{E} \left[\sum_{j=-\infty}^{\infty} B_i^j \hat{Z}_j \left(\bar{H}^{1/2} \hat{\mathbf{Z}} \right)_{\tau} \right]^2 \\
&= \frac{1}{\gamma(0)} \sum_{\tau=1}^{\infty} \sum_{i=1}^N \left(\sum_{j,l=-\infty}^{\infty} B_i^j \bar{H}_{\tau,l}^{1/2} \mathbb{E} \left[\hat{Z}_j \hat{Z}_l \right] \right)^2 = \frac{1}{\gamma(0)} \sum_{\tau=1}^{\infty} \sum_{i=1}^N \left(\sum_{j=-\infty}^{\infty} B_i^j \bar{H}_{\tau,j}^{1/2} \right)^2 \\
&= \frac{1}{\gamma(0)} \sum_{\tau=1}^{\infty} \sum_{i=1}^N \left(\bar{H}^{1/2} \mathbf{B}_i \right)_{\tau}^2 = \frac{1}{\gamma(0)} \sum_{i=1}^N \left\| \mathbb{P}_{\mathbb{Z}^+} \left(\bar{H}^{1/2} \mathbf{B}_i \right) \right\|_{\ell^2}. \quad \blacksquare
\end{aligned}$$

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 $\mathbf{Z} : \Omega \rightarrow D^{\mathbb{Z}}$ is a white noise, that is, the autocovariance function γ satisfies that $\gamma(h) = 0$, for any non-zero $h \in \mathbb{Z}$. Then:*

$$\text{MC} = N \quad \text{and} \quad \text{FC} = 0. \quad (4.8)$$

Proof. The first equality in (4.8) is a straightforward consequence of (4.3) and of the fact that for white noise inputs $H = \gamma(0) \mathbb{1}_{\ell^2_{+}(\mathbb{R})}$. The equality $\text{FC} = 0$ follows from the fact that $\bar{H} = \gamma(0) \mathbb{1}_{\ell^2(\mathbb{R})}$ and that $B_i^j = 0$, for any $i \in \{1, \dots, N\}$ and any $j \in \mathbb{Z}^+$, by (4.4). Then, by (4.5),

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

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 $A \in \mathbb{M}_N$ (usually referred to as **connectivity matrix**) and $\mathbf{C} \in \mathbb{R}^N$ in (4.1). In particular, the following proposition establishes a connection between the invertibility of $\Gamma_{\mathbf{X}}$ and Kalman's characterization of the controllability of a linear system [Kalm 10]. The relation between this condition and maximum capacity is already mentioned in [Jaeg 02].

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

- (i) *The state covariance matrix $\Gamma_{\mathbf{X}} = \text{Cov}(\mathbf{X}_t, \mathbf{X}_t) = \gamma(0) \sum_{j=0}^{\infty} A^j \mathbf{C} \mathbf{C}^{\top} (A^j)^{\top}$ is non-singular.*
- (ii) *The eigenvalues in $\sigma(A)$ are all non-zero and distinct and in the linear decomposition $\mathbf{C} = \sum_{i=1}^N c_i \mathbf{v}_i$ all the coefficients c_i , $i \in \{1, \dots, N\}$ are non-zero.*
- (iii) *The vectors $\{\mathbf{A} \mathbf{C}, \mathbf{A}^2 \mathbf{C}, \dots, \mathbf{A}^N \mathbf{C}\}$ form a basis of \mathbb{R}^N .*
- (iv) **Kalman controllability condition:** *the vectors $\{\mathbf{C}, \mathbf{A} \mathbf{C}, \dots, \mathbf{A}^{N-1} \mathbf{C}\}$ form a basis of \mathbb{R}^N .*

Proof. (i) \Leftrightarrow (ii) Using the notation introduced in the statement notice that:

$$A\mathbf{C} = \sum_{i=1}^N c_i \lambda_i \mathbf{v}_i, \quad A^2\mathbf{C} = \sum_{i=1}^N c_i \lambda_i^2 \mathbf{v}_i, \quad \dots, \quad A^N\mathbf{C} = \sum_{i=1}^N c_i \lambda_i^N \mathbf{v}_i. \quad (4.9)$$

Since by hypothesis $\|A\|_2 = \sigma_{\max}(A) < 1$, the spectral radius $\rho(A)$ of A satisfies that $\rho(A) \leq \sigma_{\max}(A) < 1$, and hence:

$$B := \sum_{j=0}^{\infty} A^j \mathbf{C} \mathbf{C}^\top (A^j)^\top = \sum_{k=0}^{\infty} \sum_{i,j=1}^N \lambda_i^k \lambda_j^k c_i c_j \mathbf{v}_i \mathbf{v}_j^\top = \sum_{i,j=1}^N \frac{c_i c_j}{1 - \lambda_i \lambda_j} \mathbf{v}_i \mathbf{v}_j^\top,$$

which shows that in the matrix basis $\{\mathbf{v}_i \mathbf{v}_j^\top\}_{i,j \in \{1, \dots, N\}}$, the matrix B has components $\bar{B}_{ij} := \frac{c_i c_j}{1 - \lambda_i \lambda_j}$ or, equivalently,

$$\bar{B} := \bar{\mathbf{C}} \bar{\mathbf{C}}^\top \odot D, \quad \text{with } \bar{\mathbf{C}} = (c_1, \dots, c_N)^\top, \quad \text{and } D \text{ defined by } D_{ij} := \frac{1}{1 - \lambda_i \lambda_j}, \quad (4.10)$$

for any $i, j \in \{1, \dots, N\}$; the symbol \odot stands for componentwise matrix multiplication (Hadamard product). Let $P \in \mathbb{M}_N$ be the invertible change-of-basis matrix between $\{\mathbf{v}_1, \dots, \mathbf{v}_N\}$ and the canonical basis $\{\mathbf{e}_1, \dots, \mathbf{e}_N\}$, that is, for any $i \in \{1, \dots, N\}$, we have that $\mathbf{v}_i = \sum_{k=1}^N P_{ik} \mathbf{e}_k$. It is easy to see that $B = P^\top \bar{B} P$ and hence the invertibility of B (and hence of $\Gamma_{\mathbf{X}}$) is equivalent to the invertibility of \bar{B} in (4.10), which we now characterize.

In order to provide an alternative expression for \bar{B} , recall that for any two vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^N$, we can write $\mathbf{v} \odot \mathbf{w} = \text{diag}(\mathbf{v})\mathbf{w}$, where $\text{diag}(\mathbf{v}) \in \mathbb{M}_N$ 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 [Horn 94, Lemma 5.1.4, page 305]) we have that

$$\begin{aligned} \langle \mathbf{v}, \bar{B} \mathbf{w} \rangle &= \left\langle \mathbf{v}, (\bar{\mathbf{C}} \bar{\mathbf{C}}^\top \odot D) \mathbf{w} \right\rangle = \text{trace} \left(\mathbf{v}^\top (\bar{\mathbf{C}} \bar{\mathbf{C}}^\top \odot D) \mathbf{w} \right) = \text{trace} \left((\bar{\mathbf{C}} \bar{\mathbf{C}}^\top \odot D) \mathbf{w} \mathbf{v}^\top \right) \\ &= \text{trace} \left((\bar{\mathbf{C}} \bar{\mathbf{C}}^\top \odot \mathbf{v} \mathbf{w}^\top) D^\top \right) = \text{trace} \left((\bar{\mathbf{C}} \odot \mathbf{v}) (\bar{\mathbf{C}} \odot \mathbf{w})^\top D^\top \right) \\ &= \text{trace} \left(\text{diag}(\bar{\mathbf{C}}) \mathbf{v} \mathbf{w}^\top \text{diag}(\bar{\mathbf{C}}) D^\top \right) = \langle \mathbf{v}, \text{diag}(\bar{\mathbf{C}}) D \text{diag}(\bar{\mathbf{C}}) \mathbf{w} \rangle. \end{aligned}$$

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

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

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) \Leftrightarrow (iii) The condition on the vectors $\{A\mathbf{C}, A^2\mathbf{C}, \dots, A^N\mathbf{C}\}$ forming a basis of \mathbb{R}^N is equivalent to the invertibility of the matrix $\widehat{R(A, \mathbf{C})} := (A\mathbf{C} | A^2\mathbf{C} | \dots | A^N\mathbf{C})$. It is easy to see using (4.9) that

$$\widehat{R(A, \mathbf{C})} = P^\top \widehat{R(A, \bar{\mathbf{C}})}, \quad (4.11)$$

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

$$\overline{R(A, \mathbf{C})} := \begin{pmatrix} c_1 \lambda_1 & c_1 \lambda_1^2 & \cdots & c_1 \lambda_1^N \\ c_2 \lambda_2 & c_2 \lambda_2^2 & \cdots & c_2 \lambda_2^N \\ \vdots & \vdots & \ddots & \vdots \\ c_N \lambda_N & c_N \lambda_N^2 & \cdots & c_N \lambda_N^N \end{pmatrix}. \quad (4.12)$$

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

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

which proves (4.11). 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) \Leftrightarrow (iv) The Kalman controllability condition on the vectors $\{\mathbf{C}, A\mathbf{C}, \dots, A^{N-1}\mathbf{C}\}$ forming a basis of \mathbb{R}^N is equivalent to the invertibility of the **controllability** or **reachability** matrix $R(A, \mathbf{C}) := (\mathbf{C} | A^2 \mathbf{C} | \cdots | A^{N-1} \mathbf{C})$ (see [Sont 98] for this terminology). Following the same strategy that we used to prove (4.11), 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 . ■

4.2 The singular case

In the following paragraphs we study the situation in which the covariance matrix of the outputs $\Gamma_{\mathbf{X}}$ in the presence of 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. The strategy that we adopt to tackle this problem uses Lemma 3.3. 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 result, *coincides with the rank of the controllability matrix $R(A, \mathbf{C})$* .

Theorem 4.4 Consider the linear system $F(\mathbf{x}, z) := \mathbf{A}\mathbf{x} + \mathbf{C}z$ introduced in (4.1) and suppose that the input process $\mathbf{Z} : \Omega \rightarrow D^{\mathbb{Z}}$ is a white noise.

(i) Let $R(\mathbf{A}, \mathbf{C})$ be the controllability matrix of the linear system. Then:

$$\ker \Gamma_{\mathbf{X}} = \ker R(\mathbf{A}, \mathbf{C})^{\top}. \quad (4.13)$$

(ii) Let $X := \text{span} \{ \mathbf{C}, \mathbf{A}\mathbf{C}, \dots, \mathbf{A}^{N-1}\mathbf{C} \}$ and let $r := \dim(X) = \text{rank} R(\mathbf{A}, \mathbf{C})$. Let $V \subset \mathbb{R}^N$ be such that $\mathbb{R}^N = X \oplus V$ and let $i_X : X \rightarrow \mathbb{R}^N$ and $\pi_X : \mathbb{R}^N \rightarrow X$ be the injection and the projection associated to this splitting, respectively. Then, the linear system $\bar{F} : X \times D \rightarrow X$ defined by

$$\bar{F}(\bar{\mathbf{x}}, z) := \bar{\mathbf{A}}\bar{\mathbf{x}} + \bar{\mathbf{C}}z \text{ and determined by } \bar{\mathbf{A}} := \pi_X \mathbf{A} i_X \text{ and } \bar{\mathbf{C}} = \pi_X(\mathbf{C}), \quad (4.14)$$

is well-defined and has the echo state property.

(iii) The inclusion $i_X : X \rightarrow \mathbb{R}^N$ is an injective linear system equivariant map between \bar{F} and F .

(iv) Let $\bar{\mathbf{X}} : \Omega \rightarrow V^{\mathbb{Z}}$ be the output of the filter determined by the state-system \bar{F} . Then:

$$\text{rank} R(\bar{\mathbf{A}}, \bar{\mathbf{C}}) = r \quad \text{and if } \bar{\mathbf{A}} \text{ is diagonalizable then } \Gamma_{\bar{\mathbf{X}}} := \text{Cov}(\bar{\mathbf{X}}_t, \bar{\mathbf{X}}_t) \text{ is invertible.} \quad (4.15)$$

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

$$\text{MC} = \text{rank} R(\mathbf{A}, \mathbf{C}) = \dim(\text{span} \{ \mathbf{C}, \mathbf{A}\mathbf{C}, \dots, \mathbf{A}^{N-1}\mathbf{C} \}) \quad \text{and} \quad \text{FC} = 0. \quad (4.16)$$

Proof. (i) We first show that $\ker R(\mathbf{A}, \mathbf{C})^{\top} \subset \ker \Gamma_{\mathbf{X}}$. Let $\mathbf{v} \in \ker R(\mathbf{A}, \mathbf{C})^{\top}$ and show that $\mathbf{C}^{\top}(\mathbf{A}^j)^{\top}\mathbf{v} = \mathbf{0}$, for all $j \in \mathbb{N}$. Indeed, the condition $\mathbf{v} \in \ker R(\mathbf{A}, \mathbf{C})^{\top}$ implies that

$$\mathbf{C}^{\top}(\mathbf{A}^j)^{\top}\mathbf{v} = \mathbf{0}, \text{ for all } j \in \{0, \dots, N-1\}. \quad (4.17)$$

Now, by the Hamilton-Cayley Theorem [Horn 13, Theorem 2.4.3.2], for any $j \geq N$, there exist constants $\{ \beta_0^j, \dots, \beta_{N-1}^j \}$ such that $\mathbf{A}^j = \sum_{i=0}^{N-1} \beta_i^j \mathbf{A}^i$ which, together with (4.17), implies that that equality holds for all $j \in \mathbb{N}$. Recall now that by Proposition 4.3 part (i), $\Gamma_{\mathbf{X}} = \gamma(0) \sum_{j=0}^{\infty} \mathbf{A}^j \mathbf{C} \mathbf{C}^{\top} (\mathbf{A}^j)^{\top}$, and hence we can conclude that $\Gamma_{\mathbf{X}}(\mathbf{v}) = \gamma(0) \sum_{j=0}^{\infty} \mathbf{A}^j \mathbf{C} \mathbf{C}^{\top} (\mathbf{A}^j)^{\top} \mathbf{v} = \mathbf{0}$, that is, $\mathbf{v} \in \ker \Gamma_{\mathbf{X}}$. Conversely, if $\mathbf{v} \in \ker \Gamma_{\mathbf{X}}$, we have that $0 = \langle \mathbf{v}, \Gamma_{\mathbf{X}}(\mathbf{v}) \rangle = \gamma(0) \sum_{j=0}^{\infty} \|\mathbf{C}^{\top} (\mathbf{A}^j)^{\top} \mathbf{v}\|^2$, which implies that $\mathbf{C}^{\top} (\mathbf{A}^j)^{\top} \mathbf{v} = \mathbf{0}$, necessarily, for any $j \in \mathbb{N}$ and hence $\mathbf{v} \in \ker R(\mathbf{A}, \mathbf{C})^{\top}$.

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

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

which implies that $\|\bar{\mathbf{A}}\|_2 \leq \|\mathbf{A}\|_2 = \sigma_{\max}(\mathbf{A}) < 1$.

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

$$\begin{aligned} \mathbf{A} i_X \bar{\mathbf{x}} + \mathbf{C}z &= \sum_{i=0}^{N-1} \alpha_i \mathbf{A}^{i+1} \mathbf{C} + \mathbf{C}z = \mathbf{C}z + \sum_{i=1}^{N-1} \alpha_{i-1} \mathbf{A}^i \mathbf{C} + \alpha_{N-1} \mathbf{A}^N \mathbf{C} \\ &= \mathbf{C}z + \sum_{i=1}^{N-1} \alpha_{i-1} \mathbf{A}^i \mathbf{C} + \alpha_{N-1} \sum_{j=0}^{N-1} \beta_j^N \mathbf{A}^j \mathbf{C} \in X, \end{aligned} \quad (4.18)$$

where the constants $\{\beta_0^N, \dots, \beta_{N-1}^N\}$ 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 [Horn 13, Theorem 2.4.3.2]. We now show that i_X is a system equivariant map between \bar{F} and F . For any $\bar{\mathbf{x}} \in X$ and $z \in D$,

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

where the second equality holds because $i_X \circ \pi_X|_X = \mathbb{I}_X$ and, by (4.18), $A i_X \bar{\mathbf{x}} + \mathbf{C}z \in X$.

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

$$i_X \left(\text{span} \left\{ \bar{\mathbf{C}}, \bar{A}\bar{\mathbf{C}}, \dots, \bar{A}^r \bar{\mathbf{C}} \right\} \right) = X.$$

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

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

which shows that $i_X \left(\text{span} \left\{ \bar{\mathbf{C}}, \bar{A}\bar{\mathbf{C}}, \dots, \bar{A}^r \bar{\mathbf{C}} \right\} \right) \subset X$. 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,

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

which clearly belongs to $i_X \left(\text{span} \left\{ \bar{\mathbf{C}}, \bar{A}\bar{\mathbf{C}}, \dots, \bar{A}^r \bar{\mathbf{C}} \right\} \right)$. The constants $\beta_{k_j}^j$ are obtained, again, by using the Hamilton-Cayley Theorem. Finally, the invertibility of $\Gamma_{\bar{\mathbf{x}}}$ is a consequence of Proposition 4.3.

(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 \bar{F} with respect to \mathbf{Z} . Now, the statement in part (iv), Proposition 4.3, and Corollary 4.2 imply that those capacities coincide with r and 0, respectively. Finally, as $r = \text{rank } R(A, \mathbf{C})$, the claim follows. ■

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 those capacities in terms of the dimensionality of the network and the autocovariance of the inputs that generalize those formulated in [Jaeg 02] for independent inputs.

The approach followed in the paper is particularly advantageous for linear networks for which explicit expressions can be formulated for the memory and the forecasting 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 [Jaeg 02]. 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.

Acknowledgments: Lukas G 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.

References

- [Aci 17] P. V. Aceituno, Y. Gang, and Y.-Y. Liu. “Tailoring artificial neural networks for optimal learning”. jul 2017.
- [Bai 12] Bai Zhang, D. J. Miller, and Yue Wang. “Nonlinear system modeling with random matrices: echo state networks revisited”. *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 23, No. 1, pp. 175–182, jan 2012.
- [Bara 14] P. Barancok and I. Farkas. “Memory capacity of input-driven echo state networks at the edge of chaos”. In: *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, pp. 41–48, 2014.
- [Broc 06] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer-Verlag, 2006.
- [Bueh 06] M. Buehner and P. Young. “A tighter bound for the echo state property”. *IEEE Transactions on Neural Networks*, Vol. 17, No. 3, pp. 820–824, 2006.
- [Char 17] A. S. Charles, D. Yin, and C. J. Rozell. “Distributed sequence memory of multidimensional inputs in recurrent networks”. Tech. Rep., 2017.
- [Coui 16] R. Couillet, G. Wainrib, H. Sevi, and H. T. Ali. “The asymptotic performance of linear echo state neural networks”. *Journal of Machine Learning Research*, Vol. 17, No. 178, pp. 1–35, 2016.
- [Damb 12] J. Dambre, D. Verstraeten, B. Schrauwen, and S. Massar. “Information processing capacity of dynamical systems”. *Scientific reports*, Vol. 2, No. 514, 2012.
- [Fark 16] I. Farkas, R. Bosak, and P. Gergel. “Computational analysis of memory capacity in echo state networks”. *Neural Networks*, Vol. 83, pp. 109–120, 2016.
- [Gall 17] C. Gallicchio and A. Micheli. “Echo state property of deep reservoir computing networks”. *Cognitive Computation*, Vol. 9, 2017.
- [Gang 08] S. Ganguli, D. Huh, and H. Sompolinsky. “Memory traces in dynamical systems.”. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 105, No. 48, pp. 18970–5, dec 2008.
- [Gono 19] L. Gonon, L. Grigoryeva, and J.-P. Ortega. “Risk bounds for reservoir computing”. *Preprint*, 2019.

- [Goud 16] A. Goudarzi, S. Marzen, P. Banda, G. Feldman, M. R. Lakin, C. Teuscher, and D. Stefanovic. “Memory and information processing in recurrent neural networks”. Tech. Rep., 2016.
- [Gray 06] R. M. Gray. “Toeplitz and Circulant Matrices: A Review”. *Foundations and Trends in Communications and Information Theory*, Vol. 2, No. 3, pp. 155–239, 2006.
- [Grig 14] L. Grigoryeva, J. Henriques, L. Larger, and J.-P. Ortega. “Stochastic time series forecasting using time-delay reservoir computers: performance and universality”. *Neural Networks*, Vol. 55, pp. 59–71, 2014.
- [Grig 15] L. Grigoryeva, J. Henriques, L. Larger, and J.-P. Ortega. “Optimal nonlinear information processing capacity in delay-based reservoir computers”. *Scientific Reports*, Vol. 5, No. 12858, pp. 1–11, 2015.
- [Grig 16a] L. Grigoryeva, J. Henriques, L. Larger, and J.-P. Ortega. “Nonlinear memory capacity of parallel time-delay reservoir computers in the processing of multidimensional signals”. *Neural Computation*, Vol. 28, pp. 1411–1451, 2016.
- [Grig 16b] L. Grigoryeva, J. Henriques, and J.-P. Ortega. “Reservoir computing: information processing of stationary signals”. In: *Proceedings of the 19th IEEE International Conference on Computational Science and Engineering*, pp. 496–503, 2016.
- [Grig 18] L. Grigoryeva and J.-P. Ortega. “Echo state networks are universal”. *Neural Networks*, Vol. 108, pp. 495–508, 2018.
- [Grig 19] L. Grigoryeva and J.-P. Ortega. “Differentiable reservoir computing”. *Journal of Machine Learning Research*, Vol. 20, No. 179, pp. 1–62, 2019.
- [Herm 10] M. Hermans and B. Schrauwen. “Memory in linear recurrent neural networks in continuous time.”. *Neural networks : the official journal of the International Neural Network Society*, Vol. 23, No. 3, pp. 341–55, apr 2010.
- [Horn 13] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, second Ed., 2013.
- [Horn 94] R. A. Horn and C. R. Johnson. *Topics in matrix analysis*. Cambridge University Press, Cambridge, 1994.
- [Jaeg 02] H. Jaeger. “Short term memory in echo state networks”. *Fraunhofer Institute for Autonomous Intelligent Systems. Technical Report.*, Vol. 152, 2002.
- [Jaeg 04] H. Jaeger and H. Haas. “Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication”. *Science*, Vol. 304, No. 5667, pp. 78–80, 2004.
- [Jaeg 10] 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.
- [Kall 02] O. Kallenberg. *Foundations of Modern Probability. Probability and Its Applications*, Springer New York, 2002.
- [Kalm 10] R. Kalman. “Lectures on Controllability and Observability”. In: *Controllability and Observability*, pp. 1–149, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

- [Livi 16] L. Livi, F. M. Bianchi, and C. Alippi. “Determination of the edge of criticality in echo state networks through Fisher information maximization”. 2016.
- [Manj 13] G. Manjunath and H. Jaeger. “Echo state property linked to an input: exploring a fundamental characteristic of recurrent neural networks”. *Neural Computation*, Vol. 25, No. 3, pp. 671–696, 2013.
- [Marz 17] S. Marzen. “Difference between memory and prediction in linear recurrent networks”. *Physical Review E*, Vol. 96, No. 3, pp. 1–7, 2017.
- [Matt 92] M. B. Matthews. *On the Uniform Approximation of Nonlinear Discrete-Time Fading-Memory Systems Using Neural Network Models*. PhD thesis, ETH Zürich, 1992.
- [Matt 93] M. B. Matthews. “Approximating nonlinear fading-memory operators using neural network models”. *Circuits, Systems, and Signal Processing*, Vol. 12, No. 2, pp. 279–307, jun 1993.
- [Matt 94] M. Matthews and G. Moschytz. “The identification of nonlinear discrete-time fading-memory systems using neural network models”. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, Vol. 41, No. 11, pp. 740–751, 1994.
- [Orti 12] S. Ortín, L. Pesquera, and J. M. Gutiérrez. “Memory and nonlinear mapping in reservoir computing with two uncoupled nonlinear delay nodes”. In: T. Gilbert, M. Kirkilionis, and G. Nicolis, Eds., *Proceedings of the European Conference on Complex Systems*, pp. 895–899, Springer International Publishing Switzerland, 2012.
- [Orti 19] S. Ortín and L. Pesquera. “Tackling the trade-off between information processing capacity and rate in delay-based reservoir computers”. *Frontiers in Physics*, Vol. 7, p. 210, 2019.
- [Orti 20] S. Ortín and L. Pesquera. “Delay-based reservoir computing: tackling performance degradation due to system response time”. *Optics Letters*, Vol. 45, No. 4, pp. 905–908, 2020.
- [Sont 91] E. D. Sontag. “Kalman’s controllability rank condition: from linear to nonlinear”. In: A. C. Antoulas, Ed., *Mathematical System Theory*, pp. 453–462, Springer, 1991.
- [Sont 98] E. Sontag. *Mathematical Control Theory: Deterministic Finite Dimensional Systems*. Springer-Verlag, 1998.
- [Tino 13] P. Tino and A. Rodan. “Short term memory in input-driven linear dynamical systems”. *Neurocomputing*, Vol. 112, pp. 58–63, 2013.
- [Tino 18] P. Tino. “Asymptotic Fisher memory of randomized linear symmetric Echo State Networks”. *Neurocomputing*, Vol. 298, pp. 4–8, 2018.
- [Verz 19] P. Verzelli, C. Alippi, and L. Livi. “Echo state networks with self-normalizing activations on the hyper-sphere”. *Scientific Reports*, Vol. 9, No. 1, p. 13887, dec 2019.
- [Verz 20] P. Verzelli, C. Alippi, L. Livi, and P. Tino. “Input representation in recurrent neural networks dynamics”. mar 2020.
- [Wain 16] G. Wainrib and M. N. Galtier. “A local echo state property through the largest Lyapunov exponent”. *Neural Networks*, Vol. 76, pp. 39–45, apr 2016.
- [Whit 04] O. White, D. Lee, and H. Sompolinsky. “Short-Term Memory in Orthogonal Neural Networks”. *Physical Review Letters*, Vol. 92, No. 14, p. 148102, apr 2004.

- [Xue 17] F. Xue, Q. Li, and X. Li. “The combination of circle topology and leaky integrator neurons remarkably improves the performance of echo state network on time series prediction.”. *PloS one*, Vol. 12, No. 7, p. e0181816, 2017.
- [Yild 12] I. B. Yildiz, H. Jaeger, and S. J. Kiebel. “Re-visiting the echo state property.”. *Neural Networks*, Vol. 35, pp. 1–9, nov 2012.