Erschienen in: Journal of Machine Learning Research : JMLR ; 19 (2018). - 24

Journal of Machine Learning Research 19 (2018) 1-40 Submitted 1/18; Revised 7/18; Published 7/18

Universal discrete-time reservoir computers with stochastic inputs and linear readouts using non-homogeneous state-affine systems

Lyudmila Grigoryeva LYUDMILA.GRIGORYEVA@UNI-KONSTANZ.DE
Department of Mathematics and Statistics
Graduate School of Decision Sciences
Universität Konstanz
Germany

Juan-Pablo Ortega JUAN-PABLO.ORTEGA@UNISG.CH
Faculty of Mathematics and Statistics
Universität Sankt Gallen
Switzerland
Centre National de la Recherche Scientifique (CNRS)
France

Editor: Yoshua Bengio

Abstract

A new class of non-homogeneous state-affine systems is introduced for use in reservoir computing. Sufficient conditions are identified that guarantee first, that the associated reservoir computers with linear readouts are causal, time-invariant, and satisfy the fading memory property and second, that a subset of this class is universal in the category of fading memory filters with stochastic almost surely uniformly bounded inputs. This means that any discrete-time filter that satisfies the fading memory property with random inputs of that type can be uniformly approximated by elements in the non-homogeneous state-affine family.

Keywords: reservoir computing, universality, state-affine systems, SAS, echo state networks, ESN, echo state affine systems, machine learning, fading memory property, linear training, stochastic signal treatment

1. Introduction

A ***reservoir computer (RC)*** (Jaeger (2010), Jaeger and Haas (2004), Maass et al. (2002), Maass (2011), Crook (2007), Verstraeten et al. (2007), Lukoševičius and Jaeger (2009)) or a ***RC system*** is a specific type of recurrent neural network determined by two maps, namely a ***reservoir*** $F : \mathbb{R}^N \times \mathbb{R}^n \longrightarrow \mathbb{R}^N$, $n, N \in \mathbb{N}$, and a ***readout*** map $h : \mathbb{R}^N \to \mathbb{R}$ that under certain hypotheses transform (or filter) an infinite discrete-time input $\mathbf{z} = (\ldots, \mathbf{z}_{-1}, \mathbf{z}_0, \mathbf{z}_1, \ldots) \in (\mathbb{R}^n)^{\mathbb{Z}}$ into an output signal $\mathbf{y} \in \mathbb{R}^{\mathbb{Z}}$ of the same type using the state-space transformation given by:

$$\begin{cases} \mathbf{x}_t = F(\mathbf{x}_{t-1}, \mathbf{z}_t), & (1.1)\\ y_t = h(\mathbf{x}_t), & (1.2) \end{cases}$$

where $t \in \mathbb{Z}$ and the dimension $N \in \mathbb{N}$ of the state vectors $\mathbf{x}_t \in \mathbb{R}^N$ will be referred to as the number of virtual ***neurons*** of the system. The expressions (1.1)-(1.2) determine a nonlinear state-space system and many of its dynamical properties (stability, controlability) have been studied for decades in the literature from that point of view.

This notion of reservoir computer (also known as ***liquid state machine***) is a significant generalization of the definitions found in the literature, where the readout map h is consistently taken to be linear. In many supervised machine learning applications, the reservoir map is randomly generated (see, for instance, the echo state networks in Jaeger (2010), Jaeger and Haas (2004)) and the memoryless readout is trained so that the output matches a given ***teaching signal*** that we denote by $\mathbf{d} \in \mathbb{R}^{\mathbb{Z}}$. Two important advantages of this approach lay on the fact that they reduce the training of a dynamic task to a static problem and, moreover, if the reservoir map is *rich* enough, good performances can be indeed attained with just linear readouts that are trained via a (eventually regularized) linear regression that minimizes the Euclidean distance between the output $\mathbf{y}$ and the teaching signal $\mathbf{d}$. These features circumvent well-known difficulties in the training of generic recurrent neural networks having to do with bifurcation phenomena (Doya (1992)) and that, despite recent progress in the regularization and training of deep RNN structures (see, for instance Graves et al. (2013), Pascanu et al. (2013), Zaremba et al. (2014), and references therein), render classical gradient descent methods non-convergent.

The interest for reservoir computing in both the machine learning and the signal processing communities has strongly increased in the last years. One of the main reasons for this fact is that some RC implementations are based on the computational capacities of certain non-neural dynamical systems (Crutchfield et al. (2010)), which opens the door to physical (optical or optoelectronic) realizations that have already been built using dedicated hardware (see, for instance, Jaeger et al. (2007), Atiya and Parlos (2000), Appeltant et al. (2011), Rodan and Tino (2011), Vandoorne et al. (2011), Larger et al. (2012), Paquot et al. (2012), Brunner et al. (2013), Vandoorne et al. (2014), Vinckier et al. (2015)) and that have shown unprecedented information processing speeds.

There are two central questions that need to be addressed when designing a machine learning paradigm, namely, the ***capacity*** and the ***universality*** problems. The capacity problem concerns generically the estimation of the error that is going to be committed in the execution of a specific task. In statistical learning and in the approximation theoretical treatment of static neural networks, this estimation has taken the form of generic bounds that incorporate various architecture parameters of the system like in Pisier (1981), Jones (1992), Barron (1993), Kurkova and Sanguineti (2005). In the specific context of reservoir computing, and in dynamic learning in general, one is interested in various notions of memory capacity that have been the subject of much research (Jaeger (2002), White et al. (2004), Ganguli et al. (2008), Hermans and Schrauwen (2010), Dambre et al. (2012), Grigoryeva et al. (2015), Couillet et al. (2016), Grigoryeva et al. (2016a)).

The universality problem consists in showing that the set of input/output functionals that can be generated with a specific architecture is dense in a sufficiently rich class, like the one containing, for example, all continuous or even all measurable functionals. For classical machine learning paradigms like neural networks, this question has given rise to well-known results that show that they can be considered as universal approximators in a static and deterministic setup (see, for instance, Kolmogorov (1956), Arnold (1957), Sprecher (1965, 1996, 1997), Cybenko (1989), Hornik et al. (1989), Rüschendorf and Thomsen (1998)).

There is no general recipe that allows one to conclude the universality of a given machine learning approach. The proof strategy depends much on the specific paradigm and, more importantly, on the nature of the inputs and the outputs. In the context of reservoir computing there are several situations for which universality has been established when the inputs/outputs are deterministic. There are two features that influence significantly the level of mathematical sophistication that is needed to conclude universality: first, the compactness of the time domain under consideration and second, if one works in continuous or discrete time. In the following paragraphs we briefly review the results that have already been obtained and, in passing, we present and put in context the contributions contained in this paper.

The compactness of the time domain is crucial because, as we will see later on, universality can be obtained as a consequence of various versions of the Stone-Weierstrass Theorem, which are invariably formulated for functions defined on a compact metric space. When the time domain is compact, this property is naturally inherited by the spaces relevant in the proofs. However, when it is not, it can still be secured using functionals that satisfy a condition introduced in Boyd and Chua (1985) known as the ***fading memory property***. The distinction between continuous and discrete time inputs is justified by the availability in the continuous setup of different tools coming from functional analysis that do not exist for discrete time.

Reservoir computing universality for compact time domains is a corollary of classical results in systems theory. Indeed, in the continuous time setup, it can be established for linear systems using polynomial readouts and for bilinear systems using linear readouts (see Fliess (1976), Sussmann (1976)). In the discrete-time setup, the situation is more convoluted when the readout is linear and required the introduction in Fliess and Normand-Cyrot (1980) of the so-called (homogeneous) ***state-affine systems (SAS)*** (see also Sontag (1979a,b)). The extension of these results to continuous non-compact time intervals was carried out in Boyd and Chua (1985) for fading memory filters using exponentially stable linear RCs with polynomial readouts and their bilinear counterparts with linear readouts (see also Maass and Sontag (2000) and Maass et al. (2002, 2004, 2007)). An extension to the non-compact discrete-time setup based on the Stone-Weierstrass theorem is, to our knowledge, not available in the literature and it is one of the main contributions of this paper. This problem has only been tackled from an internal approximation point of view, which consists in uniformly approximating the reservoir and readout maps in (1.1)-(1.2) in order to obtain an approximation of the resulting filter; this strategy has been introduced in Matthews (1992, 1993) for absolutely summable systems. The proofs in those works were unfortunately based on an invalid compactness assumption. Even though corrections were proposed in Perryman (1996) and in Stubberud and Perryman (1997b), this approach yields, in the best of cases, universality only within the reservoir filter category, while we aim at proving that statement in the much larger category of fading memory filters.

The paper is structured in three sections:

- All the notation and main definitions which are used later on in the paper are provided in Section 2. Important concepts like filters, reservoir filters, and the fading memory property are discussed.

- Section 3 contains two different universality results. The first one in Subsection 3.1 shows that the entire family of fading memory RCs itself is universal, as well as the much smaller one containing all the linear reservoirs with polynomial readouts, when certain spectral restrictions are imposed on the reservoir matrices (see below for details). The second universality result is contained in Subsection 3.2 and is one of the main contributions of the paper. Here we restrict ourselves to reservoir computers with linear readouts which are closer to the type of RCs used in applications. We introduce a non-homogeneous variant of the state-affine systems in Fliess and Normand-Cyrot (1980) and identify sufficient conditions that guarantee that the associated reservoir computers with linear readouts are causal, time-invariant, and satisfy the echo state and the fading memory properties. Finally, we state a universality result for a subset of this class which is shown to be universal in the category of fading memory filters with uniformly bounded inputs.

- These universality statements are generalized to the stochastic setup for almost surely uniformly bounded inputs in Section 4. In particular, it is shown that any discrete-time filter that has the fading memory property with almost surely uniformly bounded stochastic inputs can be uniformly approximated by elements in the non-homogeneous state-affine family.

Despite some preexisting work on the uniform approximation in probability of stochastic systems with finite memory (see Perryman (1996), Perryman and Stubberud (1997), Stubberud and Perryman

(1997a)), the universality result in the stochastic setup is, to our knowledge, the first of its type in the literature and opens the door to new developments in the learning of stochastic processes and their obvious applications to forecasting (see Galtier et al. (2014)). In the deterministic setup, RC has been very successful (see, for instance, Jaeger and Haas (2004), Pathak et al. (2017, 2018)) in the learning of the attractors of various dynamical systems. This approach is used for forecasting by path continuation of synthetically learnt proxies, which has led to several orders of magnitude accuracy improvements with respect to most standard dynamical systems forecasting techniques based on Takens' Theorem (Takens (1981)). We expect that the results in this paper should lead to comparable improvements in the density forecasting of stochastic processes.

2. Notation, definitions, and preliminary discussions

Vector and matrix notations. Polynomials. A column vector is denoted by a bold lower case symbol like $\mathbf{r}$ and $\mathbf{r}^\top$ indicates its transpose. Given a vector $\mathbf{v} \in \mathbb{R}^n$, we denote its entries by v_i or v^i, depending on the context, with $i \in \{1, \dots, n\}$; we also write $\mathbf{v} = (v_i)_{i \in \{1,\dots,n\}}$. We denote by $\mathbb{M}_{n,m}$ the space of real $n \times m$ matrices with $m, n \in \mathbb{N}$. When $n = m$, we use the symbol $\mathbb{M}_n$ to refer to the space of square matrices of order n. $\mathbb{D}_n \subset \mathbb{M}_n$ is the set of diagonal matrices of order n and $\mathbb{D}$ denotes the set of diagonal matrices of any order. Given a vector $\mathbf{v} \in \mathbb{R}^n$, we denote by $\mathrm{diag}(\mathbf{v})$ the diagonal matrix in $\mathbb{M}_n$ with the elements of $\mathbf{v}$ as diagonal entries. $\mathbb{N}\mathrm{il}_n^k \subset \mathbb{M}_n$ is the set of nilpotent matrices in $\mathbb{M}_n$ of index $k \leq n$, that is, $A \in \mathbb{N}\mathrm{il}_n^k$ if and only if $A \in \mathbb{M}_n$, $A^k = 0$, and $A^l \neq 0$ for any $l < k$. $\mathbb{N}\mathrm{il}$ denotes the set of nilpotent matrices of any order and any index. Given a matrix $A \in \mathbb{M}_{n,m}$, we denote its components by A_{ij} and we write $A = (A_{ij})$, with $i \in \{1, \dots, n\}$, $j \in \{1, \dots m\}$. Given a vector $\mathbf{v} \in \mathbb{R}^n$, the symbol $\|\mathbf{v}\|$ stands for its Euclidean norm. For any $A \in \mathbb{M}_{n,m}$, $\|A\|_2$ denotes its matrix norm induced by the Euclidean norms in $\mathbb{R}^m$ and $\mathbb{R}^n$, and satisfies that $\|A\|_2 = \sigma_{\max}(A)$, with $\sigma_{\max}(A)$ the largest singular value of A (Example 5.6.6 in Horn and Johnson (2013)). $\|A\|_2$ is sometimes referred to as the spectral norm of A (Horn and Johnson (2013)).

Let V_1, V_2, W_1, W_2 be vector spaces. The symbols $V_1 \oplus V_2$ and $V_1 \otimes V_2$ denote the corresponding direct sum and tensor product vector spaces (Hungerford (1974)), respectively, of V_1 and V_2. Given any $\mathbf{v}_1 \in V_1$ and $\mathbf{v}_2 \in V_2$, the vectors $\mathbf{v}_1 \oplus \mathbf{v}_2 \in V_1 \oplus V_2$ and $\mathbf{v}_1 \otimes \mathbf{v}_2 \in V_1 \otimes V_2$ are the direct sum and the (pure) tensor product of $\mathbf{v}_1$ and $\mathbf{v}_2$, respectively. Given two linear maps $A_1 : V_1 \longrightarrow W_1$ and $A_2 : V_2 \longrightarrow W_2$, we denote by $A_1 \oplus A_2 : V_1 \oplus V_2 \longrightarrow W_1 \oplus W_2$ and $A_1 \otimes A_2 : V_1 \otimes V_2 \longrightarrow W_1 \otimes W_2$ the associated direct sum and tensor product maps, respectively, defined by $A_1 \oplus A_2 (\mathbf{v}_1 \oplus \mathbf{v}_2) := A_1(\mathbf{v}_1) \oplus A_2(\mathbf{v}_2)$ and $A_1 \otimes A_2 (\mathbf{v}_1 \otimes \mathbf{v}_2) := A_1(\mathbf{v}_1) \otimes A_2(\mathbf{v}_2)$. The matrix representation of $A_1 \oplus A_2$ is obtained by concatenating in a block diagonal matrix the matrix representations of A_1 and A_2. As to the matrix representation of $A_1 \otimes A_2$ it is obtained via the Kronecker product of the matrix representations of A_1 and A_2 (Horn and Johnson (2013)).

Given an element $\mathbf{z} \in \mathbb{R}^n$, we denote by $\mathbb{R}[\mathbf{z}]$ the real-valued multivariate polynomials on $\mathbf{z}$ with real coefficients. Analogously, $\mathrm{Pol}(\mathbb{R}^n, \mathbb{R})$ will denote the set of real-valued polynomials on $\mathbb{R}^n$. When $z \in \mathbb{R}$ and $m, n \in \mathbb{N}$, we define the set $\mathbb{M}_{m,n}[z]$ of $\mathbb{M}_{m,n}$-valued polynomials on z with coefficients in $\mathbb{M}_{m,n}$ as

$$\mathbb{M}_{m,n}[z] := \{A_0 + zA_1 + z^2 A_2 + \cdots + z^r A_r \mid r \in \mathbb{N}, A_0, A_1, A_2, \dots, A_r \in \mathbb{M}_{m,n}\}. \tag{2.1}$$

$\mathbb{N}\mathrm{il}_n^k[z] \subset \mathbb{M}_n[z]$ is the set of nilpotent $\mathbb{M}_n$-valued polynomials on z of index k, that is, $p(z) \in \mathbb{N}\mathrm{il}_n^k[z]$ whenever k is the smallest natural number for which $p(z)^k = \mathbf{0}$, for all $z \in \mathbb{R}$. $\mathbb{N}\mathrm{il}[z]$ is the set of matrix-valued nilpotent polynomials on z of any order and any index.

Sequence spaces. $\mathbb{N}$ denotes the set of natural numbers with the zero element included. $\mathbb{Z}$ (respectively, $\mathbb{Z}_+$ and $\mathbb{Z}_-$) are the integers (respectively, the positive and the negative integers). The symbol $(\mathbb{R}^n)^{\mathbb{Z}}$ denotes the set of infinite real sequences of the form $\mathbf{z} = (\dots, \mathbf{z}_{-1}, \mathbf{z}_0, \mathbf{z}_1, \dots)$, $\mathbf{z}_i \in \mathbb{R}^n$, $i \in \mathbb{Z}$; $(\mathbb{R}^n)^{\mathbb{Z}_-}$ and $(\mathbb{R}^n)^{\mathbb{Z}_+}$ are the subspaces consisting of, respectively, left and right infinite sequences:

$(\mathbb{R}^n)^{\mathbb{Z}_-} = \{\mathbf{z} = (\ldots, \mathbf{z}_{-2}, \mathbf{z}_{-1}, \mathbf{z}_0) \mid \mathbf{z}_i \in \mathbb{R}^n, i \in \mathbb{Z}_-\}$, $(\mathbb{R}^n)^{\mathbb{Z}_+} = \{\mathbf{z} = (\mathbf{z}_0, \mathbf{z}_1, \mathbf{z}_2, \ldots) \mid \mathbf{z}_i \in \mathbb{R}^n, i \in \mathbb{Z}_+\}$. Analogously, $(D_n)^{\mathbb{Z}}$, $(D_n)^{\mathbb{Z}_-}$, and $(D_n)^{\mathbb{Z}_+}$ stand for (semi-)infinite sequences with elements in the subset $D_n \subset \mathbb{R}^n$. In most cases we shall use in these infinite product spaces either the product topology (see Chapter 2 in Munkres (2014)) or the topology induced by the supremum norm $\|\mathbf{z}\|_\infty := \sup_{t\in\mathbb{Z}} \{\|\mathbf{z}_t\|\}$. The symbols $\ell^\infty(\mathbb{R}^n)$ and $\ell^\infty_\pm(\mathbb{R}^n)$ will be used to denote the Banach spaces formed by the elements in those infinite product spaces that have a finite supremum norm $\|\cdot\|_\infty$. The symbol $B_n(\mathbf{v}, M) \subset \mathbb{R}^n$, denotes the open ball of radius $M > 0$ and center $\mathbf{v} \in \mathbb{R}^n$ with respect to the Euclidean norm. The bars over sets stand for topological closures, in particular, $\overline{B_n(\mathbf{v}, M)}$ is the closed ball.

Filters. We will refer to the maps of the type $U : (D_n)^{\mathbb{Z}} \longrightarrow \mathbb{R}^{\mathbb{Z}}$ as ***filters*** or ***operators*** and to those like $H : (D_n)^{\mathbb{Z}} \longrightarrow \mathbb{R}$ (or $H : (D_n)^{\mathbb{Z}_\pm} \longrightarrow \mathbb{R}$) as ***functionals***. A filter U is called ***causal*** when for any two elements $\mathbf{z}, \mathbf{w} \in (D_n)^{\mathbb{Z}}$ that satisfy that $\mathbf{z}_\tau = \mathbf{w}_\tau$ for all $\tau \leq t$, for any given $t \in \mathbb{Z}$, we have that $U(\mathbf{z})_t = U(\mathbf{w})_t$. Let $U_\tau : (D_n)^{\mathbb{Z}} \longrightarrow (D_n)^{\mathbb{Z}}$, $\tau \in \mathbb{Z}$, be the time delay operator defined by $U_\tau(\mathbf{z})_t := \mathbf{z}_{t-\tau}$. The filter U is called ***time-invariant*** when it commutes with the time delay operator, that is, $U_\tau \circ U = U \circ U_\tau$ (in this expression, the two time delay operators U_τ have to be understood as defined in the appropriate sequence spaces). We recall (see, for instance, Boyd and Chua (1985)) that there is a bijection between causal time-invariant filters and functionals on $(D_n)^{\mathbb{Z}_-}$. Indeed, given a time-invariant filter U, we can associate to it a functional $H_U : (D_n)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ via the assignment $H_U(\mathbf{z}) := U(\mathbf{z}^e)_0$, where $\mathbf{z}^e \in (D_n)^{\mathbb{Z}}$ is an arbitrary extension of $\mathbf{z} \in (D_n)^{\mathbb{Z}_-}$ to $(D_n)^{\mathbb{Z}}$. Conversely, for any functional $H : (D_n)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$, we can define a time-invariant causal filter $U_H : (D_n)^{\mathbb{Z}} \longrightarrow \mathbb{R}^{\mathbb{Z}}$ by $U_H(\mathbf{z})_t := H((\mathbb{P}_{\mathbb{Z}_-} \circ U_{-t})(\mathbf{z}))$, where U_{-t} is the $(-t)$-time delay operator and $\mathbb{P}_{\mathbb{Z}_-} : (D_n)^{\mathbb{Z}} \longrightarrow (D_n)^{\mathbb{Z}_-}$ is the natural projection. It is easy to verify that:

$$\begin{aligned} H_{U_H} &= H, \quad \text{for any functional} \quad H : (D_n)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}, \\ U_{H_U} &= U, \quad \text{for any causal time-invariant filter} \quad U : (D_n)^{\mathbb{Z}} \longrightarrow \mathbb{R}^{\mathbb{Z}}. \end{aligned}$$

Additionally, let $H_1, H_2 : (D_n)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ and $\lambda \in \mathbb{R}$, then $U_{H_1+\lambda H_2}(\mathbf{z}) = U_{H_1}(\mathbf{z}) + \lambda U_{H_2}(\mathbf{z})$, for any $\mathbf{z} \in (\mathbb{R}^n)^{\mathbb{Z}}$. In the following pages and when the discussion will take place in a causal and time-invariant setup, we will use the term filter to denote exchangeably the associated functional and the filter itself.

Reservoir filters. Consider now the RC system determined by (1.1)–(1.2). It is worth mentioning that, unlike in those expressions, the reservoir and the readout maps are in general defined only on subsets $D_N, D'_N \subset \mathbb{R}^N$ and $D_n \subset \mathbb{R}^n$ and not on the entire Euclidean spaces $\mathbb{R}^N$ and $\mathbb{R}^n$, that is, $F : D_N \times D_n \longrightarrow D'_N$ and $h : D'_N \to \mathbb{R}$. Reservoir systems determine a filter when the following existence and uniqueness property holds: for each $\mathbf{z} \in (D_n)^{\mathbb{Z}}$ there exists a unique $\mathbf{x} \in (D_N)^{\mathbb{Z}}$ such that for each $t \in \mathbb{Z}$, the relation (1.1) holds. This condition is known in the literature as the ***echo state property*** (see Jaeger (2010), Yildiz et al. (2012)) and has deserved much attention in the context of echo state networks (Jaeger and Haas (2004), Buehner and Young (2006), Bai Zhang et al. (2012), Wainrib and Galtier (2016), Manjunath and Jaeger (2013)). The echo state property formulated for infinite (or semi-infinite) inputs guarantees that the output of the filter at any given time does not depend on initial conditions. We emphasize that this is a genuine condition that is not automatically satisfied by all RC systems.

We will denote by $U^F : (D_n)^{\mathbb{Z}} \longrightarrow (D_N)^{\mathbb{Z}}$ the filter determined by the reservoir map via (1.1), that is, $U^F(\mathbf{z})_t := \mathbf{x}_t \in D_N$, and by $U^F_h : (D_n)^{\mathbb{Z}} \longrightarrow \mathbb{R}^{\mathbb{Z}}$ the one determined by the entire reservoir system, that is, $U^F_h(\mathbf{z})_t := h\left(U^F(\mathbf{z})_t\right) = y_t$. U^F_h will be called the ***reservoir filter*** associated to the RC system (1.1)–(1.2). The filters U^F and U^F_h are causal by construction and it can also be shown that they are necessarily time-invariant (Grigoryeva and Ortega (2018)). We can hence associate to U^F_h a ***reservoir functional*** $H^F_h : (D_n)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ determined by $H^F_h := H_{U^F_h}$.

Weighted norms and the fading memory property (FMP). Let $w : \mathbb{N} \longrightarrow (0, 1]$ be a decreasing sequence with zero limit. We define the associated ***weighted norm*** $\|\cdot\|_w$ on $(\mathbb{R}^n)^{\mathbb{Z}_-}$ associated to the

weighting sequence w as the map:

$$\begin{array}{rccc} \|\cdot\|_w: & (\mathbb{R}^n)^{\mathbb{Z}_-} & \longrightarrow & \overline{\mathbb{R}^+} \\ & \mathbf{z} & \longmapsto & \|\mathbf{z}\|_w := \sup_{t\in\mathbb{Z}_-}\{\|\mathbf{z}_t w_{-t}\|\}, \end{array}$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^n$. It is worth noting that the space

$$\ell^\infty_w(\mathbb{R}^n) := \left\{\mathbf{z}\in(\mathbb{R}^n)^{\mathbb{Z}_-} \mid \|\mathbf{z}\|_w < \infty\right\}, \tag{2.2}$$

endowed with weighted norm $\|\cdot\|_w$ forms a Banach space (Grigoryeva and Ortega (2018)).

All along the paper, we will work with ***uniformly bounded*** families of sequences, both in the deterministic and the stochastic setups. The two main properties of these subspaces in relation with the weighted norms are spelled out in the following two lemmas.

Lemma 1 *Let $M>0$ and let K_M be the set of elements in $(\mathbb{R}^n)^{\mathbb{Z}_-}$ which are uniformly bounded by M, that is,*

$$K_M := \left\{\mathbf{z}\in(\mathbb{R}^n)^{\mathbb{Z}_-} \mid \|\mathbf{z}_t\|\leq M \quad \text{for all} \quad t\in\mathbb{Z}_-\right\} = \overline{B_n(\mathbf{0},M)}^{\mathbb{Z}_-}, \tag{2.3}$$

with $\overline{B_n(\mathbf{0},M)}\subset\mathbb{R}^n$ the closed ball of radius M and center $\mathbf{0}$ in $\mathbb{R}^n$ with respect to the Euclidean norm. Then, for any weighting sequence w and $\mathbf{z}\in K_M$, we have that $\|\mathbf{z}\|_w<\infty$.

Additionally, let $\lambda,\rho\in(0,1)$ and let $w, w^\rho, w^{1-\rho}$ be the weighting sequences given by $w_t:=\lambda^t$, $w_t^\rho:=\lambda^{\rho t}$, $w_t^{1-\rho}:=\lambda^{(1-\rho)t}$, $t\in\mathbb{N}$. Then, the series $\sum_{t=0}^{\infty}\|\mathbf{z}_{-t}\|w_t$ is absolutely convergent and satisfies the inequalities:

$$\sum_{t=0}^{\infty}\|\mathbf{z}_{-t}\|w_t = \sum_{t=0}^{\infty}\|\mathbf{z}_{-t}\|\lambda^t \leq \|\mathbf{z}\|_{w^{1-\rho}}\frac{1}{1-\lambda^\rho}, \tag{2.4}$$

$$\sum_{t=0}^{\infty}\|\mathbf{z}_{-t}\|w_t = \sum_{t=0}^{\infty}\|\mathbf{z}_{-t}\|\lambda^t \leq \|\mathbf{z}\|_{w^{\rho}}\frac{1}{1-\lambda^{1-\rho}}. \tag{2.5}$$

The following result is a discrete-time version of Lemma 1 in Boyd and Chua (1985) that is easily obtained by noticing that in the discrete-time setup all functions are trivially continuous if we consider the discrete topology for their domains and, moreover, all families of functions are equicontinuous. A proof is given in the appendices for the sake of completeness.

Lemma 2 *Let $M>0$ and let K_M be as in (2.3). Let $w:\mathbb{N}\longrightarrow(0,1]$ be a weighting sequence. Then K_M is a compact topological space when endowed with the relative topology inherited from the norm topology in the Banach space $(\ell^\infty_w(\mathbb{R}^n),\|\cdot\|_w)$.*

Definition 3 *Let $D_n\subset\mathbb{R}^n$ and let $H_U:(D_n)^{\mathbb{Z}_-}\longrightarrow\mathbb{R}$ be the functional associated to the causal and time-invariant filter $U:(D_n)^{\mathbb{Z}}\longrightarrow\mathbb{R}^{\mathbb{Z}}$. We say that U has the* ***fading memory property (FMP)*** *whenever there exists a weighting sequence $w:\mathbb{N}\longrightarrow(0,1]$ such that the map $H_U:((D_n)^{\mathbb{Z}_-},\|\cdot\|_w)\longrightarrow\mathbb{R}$ is continuous. This means that for any $\mathbf{z}\in(D_n)^{\mathbb{Z}_-}$ and any $\epsilon>0$, there exists a $\delta(\epsilon)>0$ such that for any $\mathbf{s}\in(D_n)^{\mathbb{Z}_-}$ that satisfies that*

$$\|\mathbf{z}-\mathbf{s}\|_w = \sup_{t\in\mathbb{Z}_-}\{\|(\mathbf{z}_t-\mathbf{s}_t)w_{-t}\|\} < \delta(\epsilon), \quad \textit{then} \quad |H_U(\mathbf{z})-H_U(\mathbf{s})|<\epsilon.$$

*If the weighting sequence w is such that $w_t=\lambda^t$, for some $\lambda\in(0,1)$ and all $t\in\mathbb{N}$, then U is said to have the λ-**exponential fading memory property**.*

Remark 4 This formulation of the fading memory property is due to Boyd and Chua (1985) and it is the key concept that allowed these authors to extend to non-compact time intervals the first filter universality results formulated in the classical works Fréchet (1910), Wiener (1958), Brilliant (1958), and George (1959), always under compactness assumptions on the input space and the time interval in which inputs are defined.

Remark 5 In the context of reservoir filters, the fading memory property is in some occasions related to the ***Lyapunov stability*** of the autonomous system associated to the reservoir map by setting the input sequence equal to zero. This connection has been made explicit, for example, for discrete-time nonlinear state-space models that are affine in their inputs, and have direct feed-through term in the output (Zang and Iglesias (2004)) or for time delay reservoirs (Grigoryeva et al. (2016b)).

Remark 6 Time-invariant fading memory filters always have the ***bounded input, bounded output (BIBO)*** property. Indeed, if for simplicity we consider functionals H_U that map the zero input to zero, that is $H_U(\mathbf{0}) = 0$, and we want bounded outputs such that $|H_U(\mathbf{z})| < k$, for a given constant $k > 0$, by Definition 3 it suffices to consider inputs $\mathbf{z} \in (\mathbb{R}^N)^{\mathbb{Z}_-}$ such that $\|\mathbf{z}\|_\infty := \sup_{t \in \mathbb{Z}_-} \{\|\mathbf{z}_t\|\} < \delta(k)$. Indeed, if H_U has the FMP with respect to a weighting sequence w, then $\|\mathbf{z}\|_w \leq \|\mathbf{z}\|_\infty < \delta(k)$ and hence $|H_U(\mathbf{z})| < k$, as required. Another important dynamical implication of the fading memory property is the ***uniqueness of steady states*** or, equivalently, the asymptotic independence of the dynamics on the initial conditions. See Theorem 6 in Boyd and Chua (1985) for details about this fact.

The following lemma, which will be used later on in the paper, spells out how the FMP depends on the weighting sequence used to define it.

Lemma 7 *Let $D_n \subset \mathbb{R}^n$ and let $H_U : (D_n)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ be the functional associated to the causal and time-invariant filter $U : (D_n)^{\mathbb{Z}} \longrightarrow (\mathbb{R})^{\mathbb{Z}}$. If H_U has the FMP with respect to a given weighting sequence w, then it also has it with respect to any other weighting sequence w' which satisfies*

$$\frac{w_t}{w'_t} < \lambda, \quad \text{for a fixed} \quad \lambda > 0 \quad \text{and for all} \quad t \in \mathbb{N}.$$

In particular, the thesis of the lemma holds when w' dominates w, that is when $\lambda = 1$.

It can be shown (see Grigoryeva and Ortega (2018)) that when in this lemma the set $(D_n)^{\mathbb{Z}_-}$ is made of uniformly bounded sequences, that is, $(D_n)^{\mathbb{Z}_-} = K_M$, with K_M as in (2.3) then, if a filter has the FMP with respect to a given weighting sequence, it necessarily has the same property with respect to any other weighting sequence.

3. Universality results in the deterministic setup

The goal of this section is identifying families of reservoir filters that are able to uniformly approximate any time-invariant, causal, and fading memory filter with deterministic inputs with any desired degree of accuracy. Such families of reservoir computers are said to be ***universal***.

The main mathematical tool that we use is the Stone-Weierstrass theorem for polynomial subalgebras of real-valued functions defined on compact metric spaces. This approach provides us with universal families of filters as long as we can prove that, roughly speaking, their elements form polynomial algebras using a product defined in the space of functionals. More specifically, if $D_n \subset \mathbb{R}^n$ and $H_{U_1}, H_{U_2} : (D_n)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ are the functionals associated to the causal and time-invariant filters $U_1, U_2 : (D_n)^{\mathbb{Z}} \longrightarrow \mathbb{R}^{\mathbb{Z}}$, we readily define their product $H_{U_1} \cdot H_{U_2} : (D_n)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ and linear combination $H_{U_1} + \lambda H_{U_2} : (D_n)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$, $\lambda \in \mathbb{R}$, as

$$\left(H_{U_1} \cdot H_{U_2}\right)(\mathbf{z}) := H_{U_1}(\mathbf{z}) \cdot H_{U_2}(\mathbf{z}), \quad \left(H_{U_1} + \lambda H_{U_2}\right)(\mathbf{z}) := H_{U_1}(\mathbf{z}) + \lambda H_{U_2}(\mathbf{z}), \quad \mathbf{z} \in (D_n)^{\mathbb{Z}_-}. \tag{3.1}$$

This section contains two different universality results. The first one shows that polynomial algebras of filters generated by reservoir systems using the operations in (3.1) that have the fading memory property and that separate points, are able to approximate any fading memory filter. Two important consequences of this result are that the entire family of fading memory RCs itself is universal, as well as the one containing all the linear reservoirs with polynomial readouts, when certain spectral restrictions are imposed on the reservoir matrices (see below for details). In the second result, we restrict ourselves to reservoir computers with linear readouts and introduce the non-homogeneous state-affine family in order to be able to obtain a similar universality statement. The linearity restriction on the readouts makes this universality statement closer to the type of RCs used in applications and to the standard notion of reservoir system that one commonly finds in the literature (see Lukoševičius and Jaeger (2009)).

The first result can be seen as a discrete-time version of the one in Boyd and Chua (1985) for continuous-time filters, while the second one is an extension to infinite time intervals of the main approximation result in Fliess and Normand-Cyrot (1980), which was originally formulated for compact time intervals using homogeneous state-affine systems.

3.1 Universality for fading memory RCs with non-linear readouts

The following statement is a direct consequence of the compactness result in Lemma 2.3 and the Stone-Weierstrass, as formulated in Theorem 7.3.1 in Dieudonné (1969). See Appendix 6.4 for a detailed proof.

All along this subsection, we work with reservoir filters with uniformly bounded inputs in a set $K_M \subset (\mathbb{R}^n)^{\mathbb{Z}_-}$, as defined in (2.3). These filters are generated by reservoir systems $F : D_N \times \overline{B_n(\mathbf{0}, M)} \longrightarrow D_N$ and $h : D_N \to \mathbb{R}$, for some $n, N \in \mathbb{N}$, $M > 0$, and $D_N \subset \mathbb{R}^N$.

Theorem 8 *Let $K_M \subset (\mathbb{R}^n)^{\mathbb{Z}_-}$ be a subset of the type defined in* (2.3)*, I an index set, and let*

$$\mathcal{R} := \{H^{F_i}_{h_i} : K_M \longrightarrow \mathbb{R} \mid h_i \in C^\infty(D_{N_i}), F_i : D_{N_i} \times \overline{B_n(\mathbf{0}, M)} \longrightarrow D_{N_i}, i \in I, N_i \in \mathbb{N}\} \tag{3.2}$$

be a set of reservoir filters defined on K_M that have the FMP with respect to a given weighted norm $\|\cdot\|_w$. Let $\mathcal{A}(\mathcal{R})$ be the polynomial algebra generated by $\mathcal{R}$, that is, the set formed by finite products and linear combinations of elements in $\mathcal{R}$ according to the operations defined in (3.1)*. If the algebra $\mathcal{A}(\mathcal{R})$ contains the constant functionals and separates the points in K_M, then any causal, time-invariant fading memory filter $H : K_M \longrightarrow \mathbb{R}$ can be uniformly approximated by elements in $\mathcal{A}(\mathcal{R})$, that is, $\mathcal{A}(\mathcal{R})$ is dense in the set $(C^0(K_M), \|\cdot\|_w)$ of real-valued continuous functions on $(K_M, \|\cdot\|_w)$. More explicitly, this implies that for any fading memory filter H and any $\epsilon > 0$, there exist a finite set of indices $\{i_1, \ldots, i_r\} \subset I$ and a polynomial $p : \mathbb{R}^r \longrightarrow \mathbb{R}$ such that*

$$\|H - H^F_h\|_\infty := \sup_{\mathbf{z} \in K_M} \{|H(\mathbf{z}) - H^F_h(\mathbf{z})|\} < \epsilon \quad \textit{with} \quad h := p(h_{i_1}, \ldots, h_{i_r}) \quad \textit{and} \quad F := (F_{i_1}, \ldots, F_{i_r}).$$

An important fact is that *the polynomial algebra $\mathcal{A}(\mathcal{R})$ generated by a set formed by fading memory reservoir filters is made of fading memory reservoir filters.* Indeed, let $h_i \in C^\infty(D_{N_i})$, $F_i : D_{N_i} \times \overline{B_n(\mathbf{0}, M)} \longrightarrow D_{N_i}$, $i \in \{1, 2\}$, and $\lambda \in \mathbb{R}$. Then, the product $H^{F_1}_{h_1} \cdot H^{F_2}_{h_2}$ and the linear combination $H^{F_1}_{h_1} + \lambda H^{F_2}_{h_2}$ filters, as they were defined in (3.1), are such that

$$H^{F_1}_{h_1} \cdot H^{F_2}_{h_2} = H^F_h, \quad \text{with} \quad h := h_1 \cdot h_2 \in C^\infty(D_{N_1} \times D_{N_2}), \tag{3.3}$$

$$H^{F_1}_{h_1} + \lambda H^{F_2}_{h_2} = H^F_{h'}, \quad \text{with} \quad h' := h_1 + \lambda h_2 \in C^\infty(D_{N_1} \times D_{N_2}), \tag{3.4}$$

and where $F : (D_{N_1} \times D_{N_2}) \times \overline{B_n(\mathbf{0}, M)} \longrightarrow (D_{N_1} \times D_{N_2})$ is given by

$$F(((\mathbf{x}_1)_t, (\mathbf{x}_2)_t), \mathbf{z}_t) := (F_1((\mathbf{x}_1)_t, \mathbf{z}_t), F_2((\mathbf{x}_2)_t, \mathbf{z}_t)), \tag{3.5}$$

for any $((\mathbf{x}_1)_t, (\mathbf{x}_2)_t) \in D_{N_1} \times D_{N_2}$, $\mathbf{z}_t \in \overline{B_n(\mathbf{0}, M)}$, and $t \in \mathbb{Z}_-$. We emphasize that the functionals H_h^F and $H_{h'}^F$ in (3.3) and (3.4) are well defined because if the reservoir maps F_1 and F_2 satisfy the echo state property then so does F. Indeed, if $\mathbf{x}_1 \in (D_{N_1})^{\mathbb{Z}}$ and $\mathbf{x}_2 \in (D_{N_2})^{\mathbb{Z}}$ are the solutions of the reservoir equation (1.1) for F_1 and F_2 associated to the input $\mathbf{z} \in K_M$, then so is $(\mathbf{x}_1, \mathbf{x}_2) \in (D_{N_1} \times D_{N_2})^{\mathbb{Z}}$, defined by $(\mathbf{x}_1, \mathbf{x}_2)_t := ((\mathbf{x}_1)_t, (\mathbf{x}_2)_t)$, for F in (3.5).

This observation has as a consequence that the set formed by *all* the RC systems that have the echo state property and the FMP with respect to a given weighted norm $\| \cdot \|_w$ form a polynomial algebra that contains the constant functions (they can be obtained by using as readouts constant elements in $C^\infty(D_{N_i})$) and separates points (take the trivial reservoir map $F(\mathbf{x}, \mathbf{z}) = \mathbf{z}$ and use the separation property of $C^\infty(D_{N_i})$ together with time-invariance). This remark and Theorem 8 yield the following corollary.

Corollary 9 *Let $K_M \subset (\mathbb{R}^n)^{\mathbb{Z}_-}$ be a subset as defined in* (2.3) *and let*

$$\mathcal{R}_w := \{H_h^F : K_M \longrightarrow \mathbb{R} \mid h \in C^\infty(D_N), F : D_N \times \overline{B_n(\mathbf{0}, M)} \longrightarrow D_N, N \in \mathbb{N}\} \tag{3.6}$$

be the set of all reservoir filters with uniformly bounded inputs in K_M and that have the FMP with respect to a given weighted norm $\| \cdot \|_w$. Then $\mathcal{R}_w$ is universal, that is, it is dense in the set $(C^0(K_M), \| \cdot \|_w)$ of real-valued continuous functions on $(K_M, \| \cdot \|_w)$.

Remark 10 The stability of reservoir filters under products and linear combinations in (3.3)-(3.4) is a feature that allows us, in Corollary 9 and in some of the results that follow later on, to identify families of reservoir filters that are able to approximate any fading memory filter. This fact is a requirement for the application of the Stone-Weierstrass theorem but does not mean that we have to carry those operations out in the construction of approximating filters, which would indeed be difficult to implement in specific applications.

According to the previous corollary, reservoir filters that have the FMP are able to approximate any time-invariant fading memory filter. We now show that actually a much smaller family of reservoirs suffices to do that, namely, certain classes of linear reservoirs with polynomial readouts. Consider the RC system determined by the expressions

$$\begin{cases} \mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{c}\mathbf{z}_t, \quad A \in \mathbb{M}_N, \mathbf{c} \in \mathbb{M}_{N,n}, & (3.7) \\ y_t = h(\mathbf{x}_t), \quad h \in \mathbb{R}[\mathbf{x}]. & (3.8) \end{cases}$$

If this system has a reservoir filter associated (the next result provides a sufficient condition for this to happen) we denote by $H_h^{A,\mathbf{c}} : K_M \longrightarrow \mathbb{R}$ the associated functional and we refer to it as the ***linear reservoir functional*** determined by $A, \mathbf{c}$, and the polynomial h. These filters exhibit the following universality property that is proved in Appendix 6.5.

Corollary 11 *Let $K_M \subset (\mathbb{R}^n)^{\mathbb{Z}_-}$ be a subset of the type defined in* (2.3) *and let $0 < \epsilon < 1$. Consider the set $\mathcal{L}_\epsilon$ formed by all the linear reservoir systems as in* (3.7)-(3.8) *determined by matrices $A \in \mathbb{M}_N$ such that $\sigma_{\max}(A) < 1 - \epsilon$. Then, the elements in $\mathcal{L}_\epsilon$ generate λ_ρ-exponential fading memory reservoir functionals, with $\lambda_\rho := (1-\epsilon)^\rho$, for any $\rho \in (0,1)$. Equivalently, $\mathcal{L}_\epsilon \subset \mathcal{R}_{w^\rho}$, with $w_t^\rho := \lambda_\rho^t$, and $\mathcal{R}_{w^\rho}$ as in* (3.6)*. These functionals can be explicitly written as:*

$$H_h^{A,\mathbf{c}}(\mathbf{z}) = h\left(\sum_{i=0}^{\infty} A^i \mathbf{c}\mathbf{z}_{-i}\right), \quad \text{for any} \quad \mathbf{z} \in K_M. \tag{3.9}$$

This family is dense in $(C^0(K_M), \| \cdot \|_{w^\rho})$.

The same universality result can be stated for the following two smaller subfamilies of $\mathcal{L}_\epsilon$:

(i) *The family $\mathcal{DL}_\epsilon \subset \mathcal{L}_\epsilon$ formed by the linear reservoir systems in $\mathcal{L}_\epsilon$ determined by diagonal matrices $A \in \mathbb{D}$ such that $\sigma_{\max}(A) < 1 - \epsilon$.*

(ii) *The family $\mathcal{NL} \subset \mathcal{L}_\epsilon$ formed by the linear reservoir systems determined by nilpotent matrices $A \in \mathbb{N}\mathrm{il}$.*

Remark 12 The elements in the family $\mathcal{NL}$ belong automatically to $\mathcal{L}_\epsilon$ because the eigenvalues of a nilpotent matrix are always zero. This implies that if a linear reservoir system is determined by a nilpotent matrix $A \in \mathbb{N}\mathrm{il}_N^k$ of index $k \leq N$, then the reservoir functional $H_h^{A,\mathbf{c}}$ is automatically well-defined and given by a finite version of (3.9), that is,

$$H_h^{A,\mathbf{c}}(\mathbf{z}) = h\left(\sum_{i=0}^{k-1} A^i \mathbf{c} \mathbf{z}_{-i}\right), \quad \text{for any} \quad \mathbf{z} \in K_M. \tag{3.10}$$

3.2 State-affine systems and universality for fading memory RCs with linear readouts

As it was explained in the introduction, the standard notion of reservoir computing that one finds in the literature concerns architectures with linear readouts. It is is particularly convenient to work with RCs that have this feature in machine learning applications since in that case the training reduces to solving a linear regression problem. That makes training feasible when there is need for a high number of neurons, as it happens in many cases. This point makes relevant the identification of families of reservoirs that are universal when the readout is restricted to be linear, which is the subject of this subsection. In order to simplify the presentation, we restrict ourselves in this case to one-dimensional input signals, that is, all along this section we set $n = 1$. The extension to the general case is straightforward, even though more convoluted to write down (see Remark 22).

Definition 13 *Let $N \in \mathbb{N}$, $\mathbf{W} \in \mathbb{R}^N$, and let $p(z) \in \mathbb{M}_N[z]$ and $q(z) \in \mathbb{M}_{N,1}[z]$ be two polynomials on the variable z with matrix coefficients, as they were introduced in* (2.1)*. The **non-homogeneous state-affine system (SAS)** associated to p, q and $\mathbf{W}$ is the reservoir system determined by the state-space transformation:*

$$\begin{cases} \mathbf{x}_t = p(z_t)\mathbf{x}_{t-1} + q(z_t), & (3.11) \\ y_t = \mathbf{W}^\top \mathbf{x}_t. & (3.12) \end{cases}$$

Our next result spells out a sufficient condition that guarantees that the SAS reservoir system (3.11)-(3.12) has the echo state property. Moreover, it provides an explicit expression for the unique causal and time-invariant solution associated to a given input.

Proposition 14 *Consider a non-homogeneous state-affine system as in* (3.11)-(3.12) *determined by polynomials p, q, and a vector $\mathbf{W}$, with inputs defined on $I^{\mathbb{Z}}$, $I := [-1, 1]$. Assume that*

$$K_1 := \max_{z \in I} \|p(z)\|_2 = \max_{z \in I} \sigma_{\max}(p(z)) < 1. \tag{3.13}$$

Then, the reservoir system (3.11)-(3.12) *has the echo state property and for each input $\mathbf{z} \in I^{\mathbb{Z}}$ it has a unique causal and time-invariant solution given by*

$$\begin{cases} \mathbf{x}_t = \displaystyle\sum_{j=0}^{\infty} \left(\prod_{k=0}^{j-1} p(z_{t-k})\right) q(z_{t-j}), & (3.14) \\ y_t = \mathbf{W}^\top \mathbf{x}_t, & (3.15) \end{cases}$$

where

$$\prod_{k=0}^{j-1} p(z_{t-k}) := p(z_t) \cdot p(z_{t-1}) \cdots p(z_{t-j+1}).$$

Let now $K_2 := \max_{z \in I} \|q(z)\|_2$. *Then,*

$$\|\mathbf{x}_t\| \leq \frac{K_2}{1-K_1}, \quad \text{for all} \quad t \in \mathbb{Z}. \tag{3.16}$$

We will denote by $U_{\mathbf{W}}^{p,q} : I^{\mathbb{Z}} \longrightarrow \mathbb{R}^{\mathbb{Z}}$ *and* $H_{\mathbf{W}}^{p,q} : I^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ *the corresponding SAS reservoir filter and SAS functional, respectively.*

The next result presents two alternative conditions that imply the hypothesis $\max_{z \in I} \|p(z)\|_2 < 1$ in the previous proposition and that are easier to verify in practice.

Lemma 15 *Let* $p(z) \in \mathbb{M}_N[z]$ *be the polynomial given by*

$$p(z) := A_0 + zA_1 + z^2 A_2 + \cdots + z^{n_1} A_{n_1}, \quad n_1 \in \mathbb{N}.$$

Suppose that $z \in I$ *and consider the following three conditions:*

(i) *There exists a constant* $0 < \lambda < 1$, *such that* $\|A_i\|_2 = \sigma_{\max}(A_i) < \lambda$, *for any* $i \in \{0, 1, \ldots, n_1\}$, *and that at the same time satisfies that* $\lambda(n_1 + 1) < 1$.

(ii) $B_p := \|A_0\|_2 + \|A_1\|_2 + \cdots + \|A_{n_1}\|_2 < 1$.

(iii) $M_p := \max_{z \in I} \|p(z)\|_2 < 1$.

Then, condition **(i)** *implies* **(ii)** *and condition* **(ii)** *implies* **(iii)**.

We emphasize that since Proposition 14 was proved using condition **(iii)** in the previous lemma then, any of the three conditions in that statement imply the echo state property for (3.14)-(3.15) and the time-invariance of the corresponding solutions. The next result shows that the same situation holds in relation with the fading memory property.

Proposition 16 *Consider a non-homogeneous state-affine system as in* (3.11)-(3.12) *determined by polynomials* p, q, *and a vector* $\mathbf{W}$, *with inputs defined on* $I^{\mathbb{Z}}$, $I := [-1, 1]$. *If the polynomial* p *satisfies any of the three conditions in Lemma 15 then the corresponding reservoir filter has the fading memory property. More specifically, if* p *satisfies condition* **(i)** *in Lemma 15, then* $H_{\mathbf{W}}^{p,q} : (I^{\mathbb{Z}_-}, \| \cdot \|_{w^\rho}) \longrightarrow \mathbb{R}$ *is continuous with* $w_t^\rho := (n_1 + 1)^{\rho t} \lambda^{\rho t}$ *and* $\rho \in (0, 1)$ *arbitrary. The same conclusion holds for conditions* **(ii)** *and* **(iii)** *with* $w_t^\rho := B_p^{\rho t}$ *and* $w_t^\rho := M_p^{\rho t}$, *respectively.*

The importance of SAS in relation to the universality problem has to do with the fact that they form a polynomial algebra which allows us, under certain conditions, to use the Stone-Weierstrass theorem to prove a density statement. Before we show that, we observe that for any two polynomials $p_1(z) \in \mathbb{M}_{N_1, M_1}[z]$ and $p_2(z) \in \mathbb{M}_{N_2, M_2}[z]$ given by

$$p_1(z) := A_0^1 + zA_1^1 + z^2 A_2^1 + \cdots + z^{n_1} A_{n_1}^1, \tag{3.17}$$

$$p_2(z) := A_0^2 + zA_1^2 + z^2 A_2^2 + \cdots + z^{n_2} A_{n_2}^2, \tag{3.18}$$

with $n_1, n_2 \in \mathbb{N}$, their direct sum and their tensor product are also polynomials in z with matrix coefficients. More explicitly, $p_1 \oplus p_2(z) \in \mathbb{M}_{N_1+N_2, M_1+M_2}[z]$ and is written as

$$p_1 \oplus p_2(z) = A_0^1 \oplus A_0^2 + zA_1^1 \oplus A_1^2 + z^2 A_2^1 \oplus A_2^2 + \cdots + z^{n_2} A_{n_2}^1 \oplus A_{n_2}^2 + z^{n_2+1} A_{n_2+1}^1 \oplus \mathbf{0} + \cdots + z^{n_1} A_{n_1}^1 \oplus \mathbf{0}, \tag{3.19}$$

where we assumed that $n_2 \leq n_1$. Analogously, their tensor product $p_1 \otimes p_2(z) \in \mathbb{M}_{N_1 \cdot N_2, M_1 \cdot M_2}[z]$ and is written as

$$p_1 \otimes p_2(z) = \sum_{i=0}^{n_1} \sum_{j=0}^{n_2} z^{i+j} A_i^1 \otimes A_j^2. \tag{3.20}$$

The next result shows that the products and the linear combinations of SAS reservoir functionals are SAS reservoir functionals. Additionally, it makes explicit the polynomials that determine the corresponding SAS reservoir systems.

Proposition 17 *Let $H^{p_1,q_1}_{\mathbf{W}_1}, H^{p_2,q_2}_{\mathbf{W}_2} : I^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ be two SAS reservoir functionals associated to two corresponding time-invariant SAS reservoir systems. Assume that the two polynomials with matrix coefficients $p_1(z) \in \mathbb{M}_{N_1}[z]$ and $p_2(z) \in \mathbb{M}_{N_2}[z]$ satisfy that $\|p_1(z)\|_2 < 1 - \epsilon$ and $\|p_2(z)\|_2 < 1 - \epsilon$ for all $z \in I := [-1, 1]$ and a given $0 < \epsilon < 1$. Then, with the notation introduced in the expressions* (3.19) *and* (3.20)*, we have that:*

(i) *For any $\lambda \in \mathbb{R}$, the linear combination of the SAS reservoir functionals $H^{p_1,q_1}_{\mathbf{W}_1} + \lambda H^{p_2,q_2}_{\mathbf{W}_2}$ is a SAS reservoir functional and:*

$$H^{p_1,q_1}_{\mathbf{W}_1} + \lambda H^{p_2,q_2}_{\mathbf{W}_2} = H^{p_1 \oplus p_2, q_1 \oplus q_2}_{\mathbf{W}_1 \oplus \lambda \mathbf{W}_2}. \tag{3.21}$$

(ii) *The product of the SAS reservoir functionals $H^{p_1,q_1}_{\mathbf{W}_1} \cdot H^{p_2,q_2}_{\mathbf{W}_2}$ is a SAS reservoir functional and:*

$$H^{p_1,q_1}_{\mathbf{W}_1} \cdot H^{p_2,q_2}_{\mathbf{W}_2} = H^{p, q_1 \oplus q_2 \oplus (q_1 \otimes q_2)}_{\mathbf{0} \oplus \mathbf{0} \oplus (\mathbf{W}_1 \otimes \mathbf{W}_2)}, \tag{3.22}$$

where $p(z) \in \mathbb{M}_{N_{12}}[z]$, $N_{12} := N_1 + N_2 + N_1 \cdot N_2$, is the polynomial with matrix coefficients in $\mathbb{M}_{N_{12}}$ whose block-matrix expression for the three summands in $\mathbb{R}^{N_1} \oplus \mathbb{R}^{N_2} \oplus \left(\mathbb{R}^{N_1} \otimes \mathbb{R}^{N_2}\right)$ is:

$$p(z) := \begin{pmatrix} p_1(z) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & p_2(z) & \mathbf{0} \\ p_1 \otimes q_2(z) & q_1 \otimes p_2(z) & p_1 \otimes p_2(z) \end{pmatrix}. \tag{3.23}$$

The expression $p_1 \otimes p_2(z) \in \mathbb{M}_{N_1 \cdot N_2}[z]$ denotes the element defined in (3.20)*. The symbol $p_1 \otimes q_2(z)$ (respectively, $q_1 \otimes p_2(z)$) denotes the matrix of the linear map from $\mathbb{R}^{N_1}$ (respectively, $\mathbb{R}^{N_2}$) to $\mathbb{R}^{N_1} \otimes \mathbb{R}^{N_2}$ that associates to any $\mathbf{v}_1 \in \mathbb{R}^{N_1}$ the element $(p_1(z)\mathbf{v}_1) \otimes q_2(z)$ (respectively, $q_1(z) \otimes (p_2(z)\mathbf{v}_2)$, with $\mathbf{v}_2 \in \mathbb{R}^{N_2}$). When all the polynomials in* (3.23) *are written in terms of monomials using the conventions that we just mentioned and we factor out the different powers of the variable z, we obtain a polynomial with matrix coefficients in $\mathbb{M}_{N_{12}}$ and with degree* $\deg(p)$ *equal to*

$$\deg(p) = \max\left\{\deg(p_1) \cdot \deg(q_2), \deg(q_1) \cdot \deg(p_2), \deg(p_1) \cdot \deg(p_2)\right\}.$$

The equalities (3.21) *and* (3.22) *show that the SAS family forms a polynomial algebra.*

Remark 18 Notice that the linear reservoir equation (3.7) is a particular case of the SAS reservoir equation (3.11) that is obtained by taking for p and q polynomials of degree zero and one, respectively. Regarding that specific case, Proposition 17 shows that linear reservoirs with linear readouts do not form a polynomial algebra. Indeed, as it can be seen in (3.22), the product of two SAS filters involves the tensor product $q_1 \otimes q_2$ which, when q_1 and q_2 come from a linear filter, it has degree two and it is hence not compatible with a linear reservoir filter.

Theorem 19 (Universality of SAS reservoir computers) *Let $I^{\mathbb{Z}_-} \subset \mathbb{R}^{\mathbb{Z}_-}$ be the subset of real uniformly bounded sequences in $I := [-1, 1]$ as in* (2.3)*, that is,*

$$I^{\mathbb{Z}_-} := \{\mathbf{z} \in \mathbb{R}^{\mathbb{Z}_-} \mid z_t \in [-1, 1], \ \textit{for all} \quad t \leq 0\},$$

and let $\mathcal{S}_\epsilon$ be the family of functionals $H^{p,q}_{\mathbf{W}} : I^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ induced by the state-affine systems defined in (3.11)-(3.12) *that satisfy that $M_p := \max_{z\in I} \|p(z)\|_2 < 1-\epsilon$ and $M_q := \max_{z\in I} \|q(z)\|_2 < 1-\epsilon$. The family $\mathcal{S}_\epsilon$ forms a polynomial subalgebra of $\mathcal{R}_{w^\rho}$ (as defined in* (3.6)*) with $w^\rho_t := (1-\epsilon)^{\rho t}$ and $\rho \in (0,1)$ arbitrary, made of fading memory reservoir filters that contains the constant functions and separates points. The subfamily $\mathcal{S}_\epsilon$ is hence dense in the set $(C^0(I^{\mathbb{Z}_-}), \|\cdot\|_{w^\rho})$ of real-valued continuous functions on $(I^{\mathbb{Z}_-}, \|\cdot\|_{w^\rho})$.*

This statement implies that any causal, time-invariant fading memory filter $H : I^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ can be uniformly approximated by elements in $\mathcal{S}_\epsilon$. More specifically, for any fading memory filter H and any $\epsilon > 0$, there exist a natural number $N \in \mathbb{N}$, polynomials $p(z) \in \mathbb{M}_N[z], q(z) \in \mathbb{M}_{N,1}[z]$ with $M_p, M_q < 1-\epsilon$, and a vector $\mathbf{W} \in \mathbb{R}^N$ such that

$$\|H - H^{p,q}_{\mathbf{W}}\|_\infty := \sup_{z\in I^{\mathbb{Z}_-}} \{|H(z) - H^{p,q}_{\mathbf{W}}(z)|\} < \epsilon.$$

The same universality result can be stated for the smaller subfamily $\mathcal{NS}_\epsilon \subset \mathcal{S}_\epsilon$ formed by SAS reservoir systems determined by nilpotent polynomials $p(z) \in \mathbb{N}\mathrm{il}[z]$.

Remark 20 As it is stated in Theorem 19, it is the condition **(iii)** in Lemma 15 that yields a universal family of SAS fading memory reservoirs. As it can deduced from its proof (available in the Appendix 6.10), the families determined by conditions **(i)** or **(ii)** in that lemma contain SAS fading memory reservoirs but they do not form a polynomial algebra. In such cases, and according to Theorem 8, it is the algebras generated by them and not the families themselves that are universal.

Remark 21 A continuous-time analog of the universality result that we just proved can be obtained using the bilinear systems considered in Section 5.3 of Boyd and Chua (1985). In discrete time, but only when the number of time steps is finite, this universal approximation property is exhibited by homogeneous state-affine systems, that is, by setting $q(z) = \mathbf{0}$ in (3.11)-(3.12) (see Fliess and Normand-Cyrot (1980)).

Remark 22 Generalization to multidimensional signals. When the input signal is defined in $I_n^{\mathbb{Z}}$, with $I_n := [-1,1]^n$, a SAS family with the same universality properties can be defined by replacing the polynomials p and q in Definition 13, by polynomials of degree r and s of the form:

$$\begin{aligned} p(\mathbf{z}) &= \sum_{\substack{i_1,\ldots,i_n\in\{0,\ldots,r\}\\ i_1+\cdots+i_n\leq r}} z_1^{i_1}\cdots z_n^{i_n} A_{i_1,\ldots,i_n}, \quad A_{i_1,\ldots,i_n} \in \mathbb{M}_N, \quad \mathbf{z} \in I_n \\ q(\mathbf{z}) &= \sum_{\substack{i_1,\ldots,i_n\in\{0,\ldots,s\}\\ i_1+\cdots+i_n\leq s}} z_1^{i_1}\cdots z_n^{i_n} B_{i_1,\ldots,i_n}, \quad B_{i_1,\ldots,i_n} \in \mathbb{M}_{N,1}, \quad \mathbf{z} \in I_n. \end{aligned}$$

Remark 23 SAS with trigonometric polynomials. An analogous construction can be carried out using trigonometric polynomials of the type:

$$\begin{aligned} p(\mathbf{z}) &= \sum_{\substack{i_1,\ldots,i_n\in\{0,\ldots,r\}\\ i_1+\cdots+i_n\leq r}} \cos\left(i_1\cdot z_1 + \cdots + i_n\cdot z_n\right) A_{i_1,\ldots,i_n}, \quad A_{i_1,\ldots,i_n} \in \mathbb{M}_N, \quad \mathbf{z} \in I_n \\ q(\mathbf{z}) &= \sum_{\substack{i_1,\ldots,i_n\in\{0,\ldots,s\}\\ i_1+\cdots+i_n\leq s}} \cos\left(i_1\cdot z_1 + \cdots + i_n\cdot z_n\right) B_{i_1,\ldots,i_n}, \quad B_{i_1,\ldots,i_n} \in \mathbb{M}_{N,1}, \quad \mathbf{z} \in I_n. \end{aligned}$$

In this case, it is easy to establish that the resulting SAS family forms a polynomial algebra by invoking Proposition 17 and by reformulating the expressions (3.19) and (3.20) using the trigonometric identity

$$\cos(\theta)\cos(\phi) = \frac{1}{2}\left(\cos(\theta-\phi) + \cos(\theta+\phi)\right).$$

Additionally, the corresponding SAS family includes the linear family and hence the point separation property can be established as in the proof of Theorem 19 in the Appendix 6.10.

4. Reservoir universality results in the stochastic setup

This section extends the previously stated deterministic universality results to a setup in which the reservoir inputs and outputs are stochastic, that is, the reservoir is not driven anymore by infinite sequences but by discrete-time stochastic processes. We emphasize that we restrict our discussion to reservoirs that are deterministic and hence the only source of randomness in the systems considered is the stochastic nature of the input.

The results that follow are mainly based on the observation that if we adopt a uniform approximation criterion and we assume that the random inputs satisfy almost surely the uniform boundedness that we used as hypothesis in Section 3, then important features like the fading memory property or universality are naturally inherited in the stochastic setup from the deterministic case. This fact is what we call the ***deterministic-stochastic transfer principle*** and it is contained in the statement of Theorem 27 below. In particular, this result can be easily applied to show that all the universal families with deterministic inputs introduced in Section 3 are also universal in the stochastic setup when the input processes considered produce paths that, up to a set of measure zero, are uniformly bounded.

The stochastic setup. All along this section we work on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. If a condition defined on this probability space holds everywhere except for a set with zero measure, we will say that the relation is true ***almost surely***. Let $\mathbf{X} : \Omega \longrightarrow B$ be a random variable with $(B, \|\cdot\|_B)$ a normed space endowed with a σ-algebra (for example, but not necessarily, its Borel σ-algebra). Let

$$\|\mathbf{X}\|_{L^\infty} := \operatorname*{ess\,sup}_{\omega\in\Omega}\{\|\mathbf{X}(\omega)\|_B\} = \inf\left\{\rho \in \overline{\mathbb{R}^+} \mid \|\mathbf{X}\|_B \leq \rho \quad \text{almost surely}\right\}, \tag{4.1}$$

We denote by $L^\infty(\Omega, B)$ the classes of B-valued almost surely equal random variables whose norms have a finite essential supremum or that, equivalently, have almost surely bounded norms, that is,

$$L^\infty(\Omega, B) := S_B / \sim_B, \tag{4.2}$$

where

$$S_B := \{\mathbf{X} : \Omega \longrightarrow B \ \text{ random variable} \mid \|\mathbf{X}\|_{L^\infty} < \infty\}, \tag{4.3}$$

and $\sim_B$ is the equivalence relation defined on S_B as follows: two random variables $\mathbf{Y}$ and $\mathbf{Z}$ with finite $\|\cdot\|_{L^\infty}$ norm are $\sim_B$-equivalent if and only if $\mathbb{P}(\{\omega \in \Omega : \mathbf{Y}(\omega) \neq \mathbf{Z}(\omega)\} = 0$. As it is customary in the literature, we will not make a distinction in what follows between the elements in S_B and the classes in the quotient $L^\infty(\Omega, B)$. Using this identification we recall, for example, that $L^\infty(\Omega, B)$ is a vector space with the operations

$$(\mathbf{X} + \lambda\mathbf{Y})(\omega) := \mathbf{X}(\omega) + \lambda\mathbf{Y}(\omega) \tag{4.4}$$

for any $\mathbf{X}, \mathbf{Y} \in L^\infty(\Omega, B)$, $\lambda \in \mathbb{R}$, $\omega \in \Omega$. Moreover, $(L^\infty(\Omega, B), \|\cdot\|_{L^\infty})$ is a normed space. We emphasize that $L^\infty(\Omega, B)$ is in general not a Banach space (see pages 42 and 46 in Ledoux and Talagrand (1991)). It can be shown that whenever B is finite dimensional or, more generally, a separable Banach space, then the space $L^\infty(\Omega, B)$ is also a Banach space (Pisier (2016)).

Given an element $\mathbf{X} \in L^\infty(\Omega, B)$, we denote by $\mathrm{E}[\mathbf{X}]$ its expectation. The following lemma collects some elementary results that will be needed later on and shows, in particular, that the expectation $\mathrm{E}[\mathbf{X}]$ as well as that of all the powers $\|\mathbf{X}\|_B^k$ of its norm are finite for all the elements $\mathbf{X} \in L^\infty(\Omega, B)$.

Lemma 24 *Let $\mathbf{X} \in L^\infty(\Omega, B)$ and let $C \in \overline{\mathbb{R}^+}$. Then:*

(i) $\|\mathbf{X}\|_B \leq \|\mathbf{X}\|_{L^\infty}$ *almost surely.*

(ii) $\|\mathbf{X}\|_{L^\infty} \leq C$ *if and only if* $\|\mathbf{X}\|_B \leq C$ *almost surely.*

(iii) $\|\mathbf{X}\|_B \leq C$ *almost surely if and only if* $\mathrm{E}\left[\|\mathbf{X}\|_B^k\right] \leq C^k$ *for any* $k \in \mathbb{N}$.

(iv) *Let $B = \mathbb{R}^n$, then the components X_i of $\mathbf{X}$, $i \in \{1, \ldots, n\}$, are such that* $\mathrm{E}[X_i] \leq \|\mathbf{X}\|_{L^\infty}$.

The first point in this lemma explains why we will refer to the elements of $L^\infty(\Omega, B)$ as ***almost surely bounded*** random variables.

Stochastic inputs and outputs. The filters that we will consider in this section have ***almost surely bounded stochastic processes*** as inputs and outputs. Recall that a discrete-time stochastic process is a map of the type:

$$\begin{array}{rccc} \mathbf{z}: & \mathbb{Z} \times \Omega & \longrightarrow & \mathbb{R}^n \\ & (t, \omega) & \longmapsto & \mathbf{z}_t(\omega), \end{array} \tag{4.5}$$

such that, for each $t \in \mathbb{Z}$, the assignment $\mathbf{z}_t : \Omega \longrightarrow \mathbb{R}^n$ is a random variable. For each $\omega \in \Omega$, we will denote by $\mathbf{z}(\omega) := \{\mathbf{z}_t(\omega) \in \mathbb{R}^n \mid t \in \mathbb{Z}\}$ the ***realization*** or the ***sample path*** of the process $\mathbf{z}$. The results that follow are presented for stochastic processes indexed by $\mathbb{Z}$ but are equally valid for $\mathbb{Z}_+$ and $\mathbb{Z}_-$.

Recall that a map of the type (4.5) is a $\mathbb{R}^n$-valued stochastic process if and only if the corresponding map $\mathbf{z} : \Omega \longrightarrow (\mathbb{R}^n)^{\mathbb{Z}}$ into path space (designated with the same symbol) is a random variable when in $(\mathbb{R}^n)^{\mathbb{Z}}$ we consider the product sigma algebra generated by cylinder sets (Chapter 1 in Comets and Meyre (2006)). Then, the space of $\mathbb{R}^n$-valued stochastic processes can be made into a vector space with the same operations as in (4.4) and we can define in this space a norm $\|\cdot\|_{L^\infty}$ using the same prescription as in (4.1) by considering $(\mathbb{R}^n)^{\mathbb{Z}}$ as a normed space with the supremum norm $\|\cdot\|_\infty$, that is,

$$\|\mathbf{z}\|_{L^\infty} := \operatorname*{ess\,sup}_{\omega \in \Omega}\{\|\mathbf{z}(\omega)\|_\infty\} = \operatorname*{ess\,sup}_{\omega \in \Omega}\left\{\sup_{t \in \mathbb{Z}}\{\|\mathbf{z}_t(\omega)\|\}\right\}. \tag{4.6}$$

The following lemma provides an alternative characterization of the norm $\|\cdot\|_{L^\infty}$ that will be very useful in the proofs of the results that follow and in which the supremum and the essential supremum have been interchanged. The last statement contained in it complements part **(ii)** of Lemma 24 for processes.

Lemma 25 *Let $\mathbf{z} : \Omega \longrightarrow (\mathbb{R}^n)^{\mathbb{Z}}$ be a stochastic process. Then,*

$$\|\mathbf{z}\|_{L^\infty} := \operatorname*{ess\,sup}_{\omega \in \Omega}\left\{\sup_{t \in \mathbb{Z}}\{\|\mathbf{z}_t(\omega)\|\}\right\} = \sup_{t \in \mathbb{Z}}\left\{\operatorname*{ess\,sup}_{\omega \in \Omega}\{\|\mathbf{z}_t(\omega)\|\}\right\}. \tag{4.7}$$

Equivalently, using the notation in (4.1),

$$\|\mathbf{z}\|_{L^\infty} := \left\|\sup_{t \in \mathbb{Z}}\{\|\mathbf{z}_t(\omega)\|\}\right\|_{L^\infty} = \sup_{t \in \mathbb{Z}}\{\|\mathbf{z}_t\|_{L^\infty}\}. \tag{4.8}$$

These equalities imply that for any non-negative real number $C \geq 0$, the process $\mathbf{z}$ satisfies that $\|\mathbf{z}\|_{L^\infty} \leq C$ if and only if $\|\mathbf{z}_t\|_{L^\infty} \leq C$ for all $t \in \mathbb{Z}$ or, equivalently, if and only if $\sup_{t \in \mathbb{Z}}\{\|\mathbf{z}_t\|_{L^\infty}\} \leq C$.

Consider now the space $L^\infty\left(\Omega, (\mathbb{R}^n)^{\mathbb{Z}}\right)$ of processes with finite $\|\cdot\|_{L^\infty}$ norm. We refer to the elements of $L^\infty\left(\Omega, (\mathbb{R}^n)^{\mathbb{Z}}\right)$ as ***almost surely bounded time series***. Additionally, consider the space $L^\infty(\Omega, \ell^\infty(\mathbb{R}^n))$ of processes whose paths are all uniformly bounded, that is, they lay in the Banach space $(\ell^\infty(\mathbb{R}^n), \|\cdot\|_\infty)$. According to the definition in (4.2), we have for both these spaces that

$$L^\infty\left(\Omega, (\mathbb{R}^n)^{\mathbb{Z}}\right) := S_{(\mathbb{R}^n)^{\mathbb{Z}}} / \sim_{(\mathbb{R}^n)^{\mathbb{Z}}}, \quad L^\infty(\Omega, \ell^\infty(\mathbb{R}^n)) := S_{\ell^\infty(\mathbb{R}^n)} / \sim_{\ell^\infty(\mathbb{R}^n)}$$

with

$$S_{(\mathbb{R}^n)^{\mathbb{Z}}} := \left\{\mathbf{z} : \mathbb{Z} \times \Omega \longrightarrow \mathbb{R}^n \ \text{ stochastic process, } \mathbf{z}(\omega) \in (\mathbb{R}^n)^{\mathbb{Z}}, \text{ for all } \omega \in \Omega \mid \|\mathbf{z}\|_{L^\infty} < \infty\right\},$$

$$S_{\ell^\infty(\mathbb{R}^n)} := \left\{ \mathbf{z} : \mathbb{Z} \times \Omega \longrightarrow \mathbb{R}^n \text{ stochastic process, } \mathbf{z}(\omega) \in \ell^\infty(\mathbb{R}^n), \text{ for all } \omega \in \Omega \mid \|\mathbf{z}\|_{L^\infty} < \infty \right\},$$

and with the almost sure equality equivalence relations $\sim_{\ell^\infty(\mathbb{R}^n)}$ and $\sim_{(\mathbb{R}^n)^{\mathbb{Z}}}$ between stochastic processes with paths in $\ell^\infty(\mathbb{R}^n)$ and $(\mathbb{R}^n)^{\mathbb{Z}}$, respectively. The following result shows that the normed spaces $L^\infty\left(\Omega, (\mathbb{R}^n)^{\mathbb{Z}}\right)$ and $L^\infty\left(\Omega, \ell^\infty(\mathbb{R}^n)\right)$ are isomorphic.

Lemma 26 *In the setup that we just introduced the inclusion* $\iota : S_{\ell^\infty(\mathbb{R}^n)} \hookrightarrow S_{(\mathbb{R}^n)^{\mathbb{Z}}}$ *is equivariant with respect to the equivalence relations* $\sim_{\ell^\infty(\mathbb{R}^n)}$ *and* $\sim_{(\mathbb{R}^n)^{\mathbb{Z}}}$ *and drops to an isomorphism of normed spaces* $\phi : \left(L^\infty\left(\Omega, (\mathbb{R}^n)^{\mathbb{Z}}\right), \|\cdot\|_{L^\infty}\right) \longrightarrow \left(L^\infty\left(\Omega, \ell^\infty(\mathbb{R}^n)\right), \|\cdot\|_{L^\infty}\right)$*. Equivalently, the following diagram commutes*

$$\begin{array}{ccc} S_{\ell^\infty(\mathbb{R}^n)} & \xhookrightarrow{\quad\iota\quad} & S_{(\mathbb{R}^n)^{\mathbb{Z}}} \\ {\scriptstyle \Pi_{\sim_{\ell^\infty(\mathbb{R}^n)}}}\big\downarrow & & \big\downarrow{\scriptstyle \Pi_{\sim_{(\mathbb{R}^n)^{\mathbb{Z}}}}} \\ L^\infty\left(\Omega, \ell^\infty(\mathbb{R}^n)\right) & \xrightarrow{\phi} & L^\infty\left(\Omega, (\mathbb{R}^n)^{\mathbb{Z}}\right), \end{array}$$

where $\Pi_{\sim_{\ell^\infty(\mathbb{R}^n)}}$ *and* $\Pi_{\sim_{(\mathbb{R}^n)^{\mathbb{Z}}}}$ *are the canonical projections.*

Let now w be a weighting sequence and let $\|\cdot\|_w$ be the associated weighted norm in $(\mathbb{R}^n)^{\mathbb{Z}_-}$. If we replace in (4.6) the ℓ^∞ norm $\|\cdot\|_\infty$ by the weighted norm $\|\cdot\|_w$, we obtain a weighted norm $\|\cdot\|_{L^\infty_w}$ in the space of processes $\mathbf{z} : \mathbb{Z}_- \times \Omega \longrightarrow \mathbb{R}^n$ indexed by $\mathbb{Z}_-$ as:

$$\|\mathbf{z}\|_{L^\infty_w} := \operatorname*{ess\,sup}_{\omega \in \Omega} \{\|\mathbf{z}(\omega)\|_w\} = \operatorname*{ess\,sup}_{\omega \in \Omega} \left\{ \sup_{t \in \mathbb{Z}_-} \{\|\mathbf{z}_t(\omega)\| w_{-t}\} \right\}. \tag{4.9}$$

We will denote by $L^\infty_w\left(\Omega, (\mathbb{R}^n)^{\mathbb{Z}_-}\right)$ the space of processes with finite $\|\cdot\|_{L^\infty_w}$ norm. A result similar to Lemma 26 shows that the normed spaces $\left(L^\infty_w\left(\Omega, (\mathbb{R}^n)^{\mathbb{Z}_-}\right), \|\cdot\|_{L^\infty_w}\right)$ and $\left(L^\infty\left(\Omega, \ell^\infty_w(\mathbb{R}^n)\right), \|\cdot\|_{L^\infty_w}\right)$ are isomorphic. Additionally, as in Lemma 25, we have that for any $\mathbf{z} \in L^\infty_w\left(\Omega, (\mathbb{R}^n)^{\mathbb{Z}_-}\right)$:

$$\|\mathbf{z}\|_{L^\infty_w} := \operatorname*{ess\,sup}_{\omega \in \Omega} \left\{ \sup_{t \in \mathbb{Z}_-} \{\|\mathbf{z}_t(\omega)\| w_{-t}\} \right\} = \sup_{t \in \mathbb{Z}_-} \left\{ \operatorname*{ess\,sup}_{\omega \in \Omega} \{\|\mathbf{z}_t(\omega)\| w_{-t}\} \right\}. \tag{4.10}$$

Deterministic filters in a stochastic setup. As we already pointed out, we consider filters U that have almost surely bounded processes as inputs and outputs. The same conventions as in the deterministic setup are used in the identification of the different signals, namely, $\mathbf{z}$ denotes the filter input process and the symbol y is reserved for the output process. Let now $D_n \subset \mathbb{R}^n$ and let $D_n^{L^\infty_{\mathbb{Z}}} \subset L^\infty(\Omega, (\mathbb{R}^n)^{\mathbb{Z}})$ be a subset formed by processes whose paths take values in D_n almost surely. In the sequel we will restrict our attention to intrinsically ***deterministic filters*** $U : D_n^{L^\infty_{\mathbb{Z}}} \longrightarrow L^\infty(\Omega, \mathbb{R}^{\mathbb{Z}})$ that are obtained by presenting almost surely bounded stochastic inputs $\mathbf{z} \in D_n^{L^\infty_{\mathbb{Z}}} \subset L^\infty(\Omega, (\mathbb{R}^n)^{\mathbb{Z}})$ to filters $U : (D_n)^{\mathbb{Z}} \longrightarrow \mathbb{R}^{\mathbb{Z}}$ similar to those introduced in the previous section, which explains why we use the same symbol for both. This is explicitly carried out by defining the output process $U(\mathbf{z}) \in L^\infty(\Omega, \mathbb{R}^{\mathbb{Z}})$ using the convention

$$(U(\mathbf{z}))(\omega) := U(\mathbf{z}(\omega)), \quad \omega \in \Omega, \tag{4.11}$$

where on the right hand side it is the filter $U : (D_n)^{\mathbb{Z}} \longrightarrow \mathbb{R}^{\mathbb{Z}}$ which is applied to the paths $\mathbf{z}(\omega) := \{\mathbf{z}_t(\omega) \in \mathbb{R}^n \mid t \in \mathbb{Z}\} \in (D_n)^{\mathbb{Z}}$ of the process $\mathbf{z}$. We call these filters deterministic because, in view of (4.11) the dependence of the image process $(U(\mathbf{z}))(\omega) \in\in L^\infty(\Omega, \mathbb{R}^{\mathbb{Z}})$ on the probability space takes place exclusively through the dependence $\mathbf{z}(\omega)$ in the input. In this section we reserve the symbol U to denote deterministic filters $U : D_n^{L^\infty_{\mathbb{Z}}} \longrightarrow L^\infty(\Omega, \mathbb{R}^{\mathbb{Z}})$. We draw attention to the fact that assuming that the filters map into almost surely bounded processes is a genuine hypothesis that needs to be verified in each specific case considered.

The concepts of causality and time-invariance are defined as in the deterministic case by replacing equalities by almost sure equalities in the corresponding identities. More explicitly, we say that the filter $U : D_n^{L^\infty_{\mathbb{Z}}} \longrightarrow L^\infty(\Omega, \mathbb{R}^{\mathbb{Z}})$ is time-invariant when for any $\tau \in \mathbb{Z}$ and any $\mathbf{z} \in D_n^{L^\infty_{\mathbb{Z}}}$, we have that

$$(U_\tau \circ U)(\mathbf{z}) = (U \circ U_\tau)(\mathbf{z}), \quad \text{almost surely.}$$

Analogously, we say that the filter is causal with stochastic inputs when for any two elements $\mathbf{z}, \mathbf{w} \in D_n^{L^\infty_{\mathbb{Z}}}$ that satisfy that $\mathbf{z}_\tau = \mathbf{w}_\tau$ almost surely, for any $\tau \leq t$ and for a given $t \in \mathbb{Z}$, we have that $U(\mathbf{z})_t = U(\mathbf{w})_t$, almost surely. Causal and time-invariant deterministic filters produce almost surely causal and time-invariant filters when stochastic inputs are presented to them.

In this setup, there is also a correspondence between causal and time-invariant filters $U : D_n^{L^\infty_{\mathbb{Z}}} \longrightarrow L^\infty(\Omega, \mathbb{R}^{\mathbb{Z}})$ and functionals $H_U : D_n^{L^\infty_{\mathbb{Z}_-}} \longrightarrow L^\infty(\Omega, \mathbb{R})$, where $D_n^{L^\infty_{\mathbb{Z}_-}} := \mathbb{P}_{\mathbb{Z}_-}\left(D_n^{L^\infty_{\mathbb{Z}}}\right)$.

Given a weighting sequence $w : \mathbb{N} \longrightarrow (0,1]$ and a time-invariant filter $U : D_n^{L^\infty_{\mathbb{Z}_-}} \longrightarrow L^\infty(\Omega, \mathbb{R}^{\mathbb{Z}})$ with stochastic inputs, we will say that U has the ***fading memory property*** with respect to the weighting sequence w when the corresponding functional $H_U : \left(D_n^{L^\infty_{\mathbb{Z}_-}}, \| \cdot \|_{L^\infty_w}\right) \longrightarrow L^\infty(\Omega, \mathbb{R})$ is a continuous map.

Let $M > 0$ and define, using Lemma 25,

$$K_M^{L^\infty} := \left\{ \mathbf{z} \in L^\infty(\Omega, (\mathbb{R}^n)^{\mathbb{Z}_-}) \mid \|\mathbf{z}\|_{L^\infty} \leq M \right\} = \left\{ \mathbf{z} \in L^\infty(\Omega, (\mathbb{R}^n)^{\mathbb{Z}_-}) \mid \|\mathbf{z}_t\|_{L^\infty} \leq M, \text{ for all } t \in \mathbb{Z}_- \right\}. \tag{4.12}$$

The sets $K_M^{L^\infty}$ are the stochastic counterparts of the sets K_M in the deterministic setup; we will say that $K_M^{L^\infty}$ is a set of ***almost surely uniformly bounded processes***. A stochastic analog of Lemma 1 can be formulated for them with K_M replaced by $K_M^{L^\infty}$, the norm $\|\cdot\|$ by $\|\cdot\|_{L^\infty}$, and the weighted norm $\|\cdot\|_w$ by $\|\cdot\|_{L^\infty_w}$. Indeed, the following result shows that the fading memory and the universality properties are naturally inherited by deterministic filters with almost surely uniformly bounded inputs. We call this fact the ***deterministic-stochastic transfer principle***.

Theorem 27 (Deterministic-stochastic transfer principle) *Let $M > 0$ and let K_M and $K_M^{L^\infty}$ be the sets of deterministic and stochastic inputs defined in (2.3) and (4.12), respectively. The following properties hold true:*

(i) *Let $H : (K_M, \|\cdot\|_w) \longrightarrow \mathbb{R}$ be a causal and time-invariant filter. Then H has the fading memory property if and only if the corresponding filter with almost surely uniformly bounded inputs has almost surely bounded outputs, that is, $H : (K_M^{L^\infty}, \| \cdot \|_{L^\infty_w}) \longrightarrow L^\infty(\Omega, \mathbb{R})$, and it has the fading memory property.*

(ii) *Let $\mathcal{T} := \{H_i : (K_M, \|\cdot\|_w) \longrightarrow \mathbb{R} \mid i \in I\}$ be a family of causal and time-invariant fading memory filters. Then, $\mathcal{T}$ is dense in the set $(C^0(K_M), \| \cdot \|_w)$ if and only if the corresponding family with inputs in $K_M^{L^\infty}$ is universal in the set of continuous maps of the type $H : (K_M^{L^\infty}, \| \cdot \|_{L^\infty_w}) \longrightarrow L^\infty(\Omega, \mathbb{R})$.*

A first universality result using RC systems. The following paragraphs contain a stochastic analog of Theorem 8 which shows that any fading memory filter with almost surely uniformly bounded inputs can be approximated using the elements of a polynomial algebra of reservoir filters with the same kind of inputs, provided that it contains the constant functionals and has the separation property. We note that, as in the deterministic case, the existence of the reservoir filter associated to a reservoir system like (1.1)-(1.2) is guaranteed only in the presence of the echo state property. The next lemma shows that this property is inherited by deterministic fading memory reservoir filters with almost surely bounded inputs.

Lemma 28 *Consider a reservoir system determined by the relations* (1.1)–(1.2) *and the maps* $F : D_N \times \overline{B_n(\mathbf{0}, M)} \longrightarrow D_N$ *and* $h : D_N \to \mathbb{R}$, *for some* $n, N \in \mathbb{N}$, $M > 0$, *and* $D_N \subset \mathbb{R}^N$. *If this reservoir system has the echo state and the fading memory properties then so does the corresponding system with stochastic inputs in* $K_M^{L^\infty}$ *which, additionally, has an associated reservoir functional* $H_h^F : (K_M^{L^\infty}, \|\cdot\|_{L_w^\infty}) \longrightarrow L^\infty(\Omega, \mathbb{R})$ *with almost surely bounded outputs that satisfies the fading memory property.*

Theorem 29 *Let* $M > 0$ *and let* $K_M^{L^\infty}$ *be the set of almost surely uniformly bounded processes introduced in* (4.12). *Consider the set* $\mathcal{R}$

$$\mathcal{R} := \{H_{h_i}^{F_i} : K_M^{L^\infty} \longrightarrow L^\infty(\Omega, \mathbb{R}) \mid h_i \in \mathrm{Pol}(\mathbb{R}^{N_i}, \mathbb{R}), F_i : \mathbb{R}^{N_i} \times \mathbb{R}^n \longrightarrow \mathbb{R}^{N_i}, i \in I, N_i \in \mathbb{N}\} \tag{4.13}$$

formed by deterministic fading memory reservoir filters with respect to a given weighted norm $\|\cdot\|_w$ *and driven by stochastic inputs in* $K_M^{L^\infty}$. *Let* $\mathcal{A}(\mathcal{R})$ *be the polynomial algebra generated by* $\mathcal{R}$. *If the algebra* $\mathcal{A}(\mathcal{R})$ *has the separation property and contains all the constant functionals, then any deterministic, causal, time-invariant fading memory filter* $H : (K_M^{L^\infty}, \|\cdot\|_{L_w^\infty}) \longrightarrow L^\infty(\Omega, \mathbb{R})$ *can be uniformly approximated by elements in* $\mathcal{A}(\mathcal{R})$, *that is, for any* $\epsilon > 0$, *there exist a finite set of indices* $\{i_1, \ldots, i_r\} \subset I$ *and a polynomial* $p : \mathbb{R}^r \longrightarrow \mathbb{R}$ *such that*

$$\|H - H_h^F\|_\infty := \sup_{\mathbf{z} \in K_M^{L^\infty}} \{\|H(\mathbf{z}) - H_h^F(\mathbf{z})\|_{L^\infty}\} < \epsilon \quad \text{with} \quad h := p(h_{i_1}, \ldots, h_{i_r}) \quad \text{and} \quad F := (F_{i_1}, \ldots, F_{i_r}).$$

In the next paragraphs we identify, as in the deterministic case, families of reservoirs that satisfy the hypotheses of this theorem and where we will eventually impose linearity constraints on the readout function. The following corollary to Theorem 29 is the stochastic analog of Corollary 9.

Corollary 30 *Let* $M > 0$ *and let* $K_M^{L^\infty}$ *be the set of almost surely uniformly bounded processes introduced in* (4.12). *Let*

$$\mathcal{R}_w := \{H_h^F : K_M^{L^\infty} \longrightarrow L^\infty(\Omega, \mathbb{R}) \mid h \in \mathrm{Pol}(\mathbb{R}^N, \mathbb{R}), F : \mathbb{R}^N \times \mathbb{R}^n \longrightarrow \mathbb{R}^N, N \in \mathbb{N}\} \tag{4.14}$$

be the set of all the reservoir filters defined on $K_M^{L^\infty}$ *that have the FMP with respect to a given weighted norm* $\|\cdot\|_{L_w^\infty}$. *Then* $\mathcal{R}_w$ *is universal, that is, for any time-invariant fading memory filter* $H : (K_M^{L^\infty}, \|\cdot\|_{L_w^\infty}) \longrightarrow L^\infty(\Omega, \mathbb{R})$ *and any* $\epsilon > 0$, *there exists a reservoir filter* $H_h^F \in \mathcal{R}_w$ *such that* $\|H - H_h^F\|_\infty := \sup_{\mathbf{z} \in K_M^{L^\infty}} \{\|H(\mathbf{z}) - H_h^F(\mathbf{z})\|_{L^\infty}\} < \epsilon$.

Linear reservoir computers with stochastic inputs are universal. As it was the case in the deterministic setup, we can prove in the stochastic case that the linear RC family introduced in (3.7)-(3.8) suffices to achieve universality. The proof of the following statement is a direct consequence of Corollary 11 and Theorem 27.

Corollary 31 *Let* $M > 0$ *and let* $K_M^{L^\infty}$ *be the set of almost surely uniformly bounded processes introduced in* (4.12). *Let* $\mathcal{L}_\epsilon$ *be the family introduced in Corollary 11 and formed by all the linear reservoir filters* $H_p^{A,\mathbf{c}}$ *determined by matrices* $A \in \mathbb{M}_N$ *such that* $\sigma_{\max}(A) < 1 - \epsilon$. *The elements in* $\mathcal{L}_\epsilon$ *map* $K_M^{L^\infty}$ *into* $L^\infty(\Omega, \mathbb{R})$ *and are time-invariant fading memory filters with respect to the weighted norm* $\|\cdot\|_{w^\rho}^{L^\infty}$ *associated to* $w_t^\rho := (1-\epsilon)^{\rho t}$, *for any* $\rho \in (0,1)$. *Moreover, they are universal, that is, for any time-invariant and causal fading memory filter* $H : (K_M^{L^\infty}, \|\cdot\|_{L_{w^\rho}^\infty}) \longrightarrow L^\infty(\Omega, \mathbb{R})$ *and any* $\varepsilon > 0$, *there exists* $H_p^{A,\mathbf{c}} \in \mathcal{L}_\epsilon$ *such that* $\|H - H_p^{A,\mathbf{c}}\|_\infty := \sup_{\mathbf{z} \in K_M^{L^\infty}} \{\|H(\mathbf{z}) - H_p^{A,\mathbf{c}}(\mathbf{z})\|_{L^\infty}\} < \varepsilon$.

The same universality result can be stated for the subfamily $\mathcal{DL}_\epsilon \subset \mathcal{L}_\epsilon$, *formed by the linear reservoir systems in* $\mathcal{L}_\epsilon$ *determined by diagonal matrices, and for* $\mathcal{NL} \subset \mathcal{L}_\epsilon$, *formed by the linear reservoir systems determined by nilpotent matrices.*

Remark 32 The linear reservoir filters in $\mathcal{NL}$ determined by nilpotent matrices have been used in Gonon and Ortega (2018) to formulate a L^p version of these universality results.

Remark 33 The previous corollary has interesting consequences in the realm of time series analysis. Indeed, many well-known parametric time series models consist in autoregressive relations, possibly nonlinear, driven by independent or uncorrelated innovations. The parameter constraints that are imposed on them in order to ensure that they have (second order) stationary solutions imply, in may situations, that the resulting filter has the FMP. In those cases, Corollary 31 allows us to conclude that when those models are driven by almost surely uniformly bounded innovations, they can be arbitrarily well approximated by a polynomial function of a vector autoregressive model (VAR) of order 1. This statement applies, for example, to any stationary ARMA (see Box and Jenkins (1976), Brockwell and Davis (2006)) or GARCH model (see Engle (1982), Bollerslev (1986), Francq and Zakoian (2010)) driven by almost surely uniformly bounded innovations.

State-affine reservoir computers with almost surely uniformly bounded inputs are universal. As it was the case in the deterministic setup, non-homogeneous SAS are universal time-invariant fading memory filters in the stochastic framework with almost surely uniformly bounded inputs. Their advantage with respect to the families proposed in the previous corollary is that they use a linear readout which is of major importance in practical implementations. More specifically, the following result holds true as a direct consequence of Theorem 19 and the equivalence stated in Theorem 27.

Theorem 34 (Universality of SAS reservoir computers with almost surely uniformly bounded inputs) *Let $K_I^{L^\infty} \subset L^\infty(\Omega, \mathbb{R}^{\mathbb{Z}_-})$ be the set of almost surely and uniformly bounded processes in the interval $I = [-1, 1]$, that is,*

$$K_I^{L^\infty} := \left\{ z \in L^\infty(\Omega, \mathbb{R}^{\mathbb{Z}_-}) \mid \|z_t\|_{L^\infty} \leq 1, \quad \text{for all} \quad t \in \mathbb{Z}_- \right\}.$$

Let $\mathcal{S}_\epsilon$ be the family of functionals $H_{\mathbf{W}}^{p,q} : K_I^{L^\infty} \longrightarrow L^\infty(\Omega, \mathbb{R})$ induced by the state-affine systems defined in (3.11)-(3.12) and that satisfy $M_p := \max_{z \in I} \|p(z)\| < 1 - \epsilon$ and $M_q := \max_{z \in I} \|q(z)\| < 1 - \epsilon$. The family $\mathcal{S}_\epsilon$ forms a polynomial subalgebra of $\mathcal{R}_{w^\rho}$ (as defined in (4.14)) with $w_t^\rho := (1-\epsilon)^{\rho t}$, made of fading memory reservoir filters that map into $L^\infty(\Omega, \mathbb{R})$.

Moreover, for any time-invariant and causal fading memory filter $H : (K_I^{L^\infty}, \|\cdot\|_{L^\infty_{w^\rho}}) \longrightarrow L^\infty(\Omega, \mathbb{R})$ and any $\epsilon > 0$, there exist a natural number $N \in \mathbb{N}$, polynomials $p(z) \in \mathbb{M}_{N,N}[z], q(z) \in \mathbb{M}_{N,1}[z]$ with $M_p, M_q < 1 - \epsilon$, and a vector $\mathbf{W} \in \mathbb{R}^N$ such that

$$\|H - H_{\mathbf{W}}^{p,q}\|_\infty := \sup_{z \in K_I^{L^\infty}} \{\|H(z) - H_{\mathbf{W}}^{p,q}(z)\|_{L^\infty}\} < \epsilon.$$

The same universality result can be stated for the smaller subfamily $\mathcal{NS}_\epsilon \subset \mathcal{S}_\epsilon$ formed by SAS reservoir systems determined by nilpotent polynomials $p(z) \in \mathbb{N}\mathrm{il}[z]$.

5. Conclusion

This paper studies and proposes solutions for the universality problem in the approximation of fading memory filters using reservoir computer (RC) systems. RCs are a particular type of recurrent neural networks that have important applications both in machine learning and in signal processing where they exhibit superb information processing performances. Their importance is also linked to the possibility of building highly efficient hardware realizations. RC systems are in general defined as nonlinear state-space systems determined by a reservoir and a readout map. In many supervised machine learning applications the readout is chosen to be linear and the reservoir map is randomly generated, which reduces the training of a dynamic task to a static regression problem and allows to circumvent well-known difficulties in the training of generic recurrent neural networks.

The universality question that we addressed consists in finding families of RCs as simple as possible such that the set of input/output functionals that can be generated with them is dense in a sufficiently rich class. The work presented here is the dynamic counterpart of a statement of this type for neural networks in a static and deterministic setup in which they have been proved to be universal approximators.

The RC universality results stated in the paper correspond to two different situations in which the inputs are either deterministic and uniformly bounded or stochastic and almost surely uniformly bounded. In both cases we proved two different universality statements. First, we showed that the family of fading memory RCs is universal in the much larger fading memory filters category. The same applies to the much smaller RC family containing just linear reservoirs with polynomial readouts, when certain spectral restrictions are imposed on the reservoir maps. The second result concerns exclusively reservoir computers with linear readouts, which are closer to the type of RCs used in applications and hardware implementations. More specifically, we introduced the family of what we called non-homogeneous state-affine systems and identified sufficient conditions that guarantee that the associated reservoir computers with linear readouts are causal, time-invariant, and satisfy the echo state and the fading memory properties. Finally, we stated a universality result for a subset of this class which was shown to be universal in the same fading memory filters category as above. These universality statements are then generalized to the stochastic setup for almost surely uniformly bounded inputs. In particular, we showed that any discrete-time filter that has the fading memory property with almost surely uniformly bounded stochastic inputs can be uniformly approximated by elements in the non-homogeneous state-affine family. All the density statements in the paper are formulated with respect to natural uniform approximation norms that appear in each of the different cases considered.

Despite preexisting work, these universality results are, to our knowledge, the first of their type in the semi-infinite discrete-time inputs setup. In the stochastic case they open the door to new developments in the learning theory of stochastic processes.

6. Appendices

6.1 Proof of Lemma 1

Let $w : \mathbb{N} \longrightarrow (0,1]$ be an arbitrary weighting sequence. Then, for any $\mathbf{z} \in K_M$:

$$\|\mathbf{z}\|_w := \sup_{t \in \mathbb{Z}_-} \{\|\mathbf{z}_t w_{-t}\|\} = \sup_{t \in \mathbb{Z}_-} \{\|\mathbf{z}_t\| w_{-t}\} \leq M \cdot 1 = M < \infty.$$

Regarding the inequalities (2.4) and (2.5), notice that if $w_t = \lambda^t$ then:

$$\sum_{t=0}^{\infty} \|\mathbf{z}_{-t}\| w_t = \sum_{t=0}^{\infty} \|\mathbf{z}_{-t}\| \lambda^t = \sum_{t=0}^{\infty} \|\mathbf{z}_{-t}\| (\lambda^{1-\rho}\lambda^{\rho})^t = \sum_{t=0}^{\infty} \|\mathbf{z}_{-t}\| \lambda^{(1-\rho)t}\lambda^{\rho t}$$
$$\leq \sum_{t=0}^{\infty} \sup_{i \in \mathbb{N}} \left\{\|\mathbf{z}_{-i}\| \lambda^{(1-\rho)i}\right\} \lambda^{\rho t} = \sup_{i \in \mathbb{N}} \left\{\|\mathbf{z}_{-i}\| \lambda^{(1-\rho)i}\right\} \sum_{t=0}^{\infty} \lambda^{\rho t} = \|\mathbf{z}\|_{w^{1-\rho}} \frac{1}{1-\lambda^{\rho}},$$

which proves (2.4). The proof of (2.5) is similar and follows from noticing that:

$$\sum_{t=0}^{\infty} \|\mathbf{z}_{-t}\| \lambda^{(1-\rho)t}\lambda^{\rho t} \leq \sum_{t=0}^{\infty} \sup_{i \in \mathbb{N}} \left\{\|\mathbf{z}_{-i}\| \lambda^{\rho i}\right\} \lambda^{(1-\rho)t)} = \sup_{i \in \mathbb{N}} \left\{\|\mathbf{z}_{-i}\| \lambda^{\rho i}\right\} \sum_{t=0}^{\infty} \lambda^{(1-\rho)t} = \|\mathbf{z}\|_{w^{\rho}} \frac{1}{1-\lambda^{1-\rho}}. \quad \blacksquare$$

6.2 Proof of Lemma 2

We recall first that by Lemma 1 we have that $\|\mathbf{z}\|_w < \infty$, for any $\mathbf{z} \in K_M$. Second, since $(\ell_w^{\infty}(\mathbb{R}^n), \|\cdot\|_w)$ is a Banach space (Grigoryeva and Ortega (2018)), it is hence metrizable and therefore so is $(K_M, \|\cdot\|_w)$ when endowed with the relative topology (see, for instance, Exercise 1, Chapter 2, §21, Munkres (2014)). We will then conclude the compactness of $(K_M, \|\cdot\|_w)$ by showing that this space is sequentially compact (see, for example, Theorem 28.2 in Munkres (2014)). We proceed by using the strategy in the proof of Lemma 1 in Boyd and Chua (1985).

For any $m \in \mathbb{N}$, let K_M^m be the set obtained by projecting into $(\mathbb{R}^n)^{\{-m,\ldots,-1,0\}}$ the elements of $K_M \subset (\mathbb{R}^n)^{\mathbb{Z}_-}$. Given an element $\mathbf{z} \in K_M$, we will denote by $\mathbf{z}^{(m)} := (\mathbf{z}_{-m}, \ldots, \mathbf{z}_0)$ its projection into K_M^m. Additionally, notice that $K_M^m = \overline{B_n(\mathbf{0}, M)}^{m+1}$ is compact (and hence sequentially compact) with the product topology, since it is a product of closed balls $\overline{B_n(\mathbf{0}, M)} \subset \mathbb{R}^n$ which are compact.

Let $\{\mathbf{z}(n)\}_{n\in\mathbb{N}} \subset K_M$ be a sequence of elements in K_M. The argument that we just stated proves that for any $k \in \mathbb{N}$, there is a subset $\mathbb{N}_k \subset \mathbb{N}$ and an element $\mathbf{z}^{(k)} \in K_M^k$ such that

$$\max_{t\in\{-k,\ldots,0\}} \left\| \mathbf{z}_t(n) - \mathbf{z}_t^{(k)} \right\| \longrightarrow 0, \quad \text{as } n \to \infty,\ n \in \mathbb{N}_k.$$

Moreover, the sets $\mathbb{N}_k$ can be constructed so that $\mathbb{N} \supset \mathbb{N}_1 \supset \mathbb{N}_2 \supset \cdots$ and so that $\mathbf{z}^{(k)}$ extends $\mathbf{z}^{(l)}$ when $k \geq l$. This implies the existence of an element $\mathbf{z} \in K_M$ such that, for each $k \in \mathbb{N}$,

$$\max_{t\in\{-k,\ldots,0\}} \|\mathbf{z}_t(n) - \mathbf{z}_t\| \longrightarrow 0, \quad \text{as } n \to \infty,\ n \in \mathbb{N}_k,$$

and hence there exists an increasing subsequence n_k such that $n_k \in \mathbb{N}_k$ and that for each k_0,

$$\max_{t\in\{-k_0,\ldots,0\}} \|\mathbf{z}_t(n_k) - \mathbf{z}_t\| \longrightarrow 0, \quad \text{as } k \longrightarrow \infty. \tag{6.1}$$

We conclude by showing that the sequence $\{\mathbf{z}(n_k)\}_{k\in\mathbb{N}}$ converges in $(K_M, \|\cdot\|_w)$ to the element $\mathbf{z} \in K_M$. First, given that $w_t \longrightarrow 0$ as $t \longrightarrow \infty$, then for any $\varepsilon > 0$ there exists k_0 such that $w_k < \varepsilon/2M$, for any $k \geq k_0$. Additionally, since $\mathbf{z}(n_k), \mathbf{z} \in K_M$ for any $k \in \mathbb{N}$, we have that

$$\sup_{t\leq -k_0} \{\|\mathbf{z}_t(n_k) - \mathbf{z}_t\| \, w_{-t}\} \leq 2M w_{k_0} < \varepsilon. \tag{6.2}$$

Now, by (6.1) there exists k_1 such that for any $k \geq k_1$

$$\sup_{t\in\{-k_0,\ldots,0\}} \{\|\mathbf{z}_t(n_k) - \mathbf{z}_t\| \, w_{-t}\} < \sup_{t\in\{-k_0,\ldots,0\}} \{\|\mathbf{z}_t(n_k) - \mathbf{z}_t\|\} < \varepsilon. \tag{6.3}$$

Consequently, (6.2) and (6.3) imply that for any $k > \max\{k_0, k_1\}$, $\|\mathbf{z}(n_k) - \mathbf{z}\|_w < \varepsilon$, as required. ∎

6.3 Proof of Lemma 7

Let $\delta^w(\epsilon)$ be the epsilon-delta relation for the FMP associated to the weighting sequence w. We now show that H_U has the FMP with respect to w' via the epsilon-delta relation given by $\delta^{w'}(\epsilon) := \delta^w(\epsilon)/\lambda$. Indeed, for any $\epsilon > 0$ and any $\mathbf{z}, \mathbf{s} \in K$ such that $\|\mathbf{z} - \mathbf{s}\|_{w'} < \delta^{w'}(\epsilon)$, we have that

$$\|\mathbf{z}-\mathbf{s}\|_w = \sup_{t\in\mathbb{Z}_-} \{\|\mathbf{z}_t - \mathbf{s}_t\| w_{-t}\} = \sup_{t\in\mathbb{Z}_-} \left\{ \|\mathbf{z}_t - \mathbf{s}_t\| \frac{w_{-t}}{w'_{-t}} w'_{-t} \right\} < \lambda \sup_{t\in\mathbb{Z}_-} \{\|\mathbf{z}_t - \mathbf{s}_t\| w'_{-t}\} < \lambda \|\mathbf{z}-\mathbf{s}\|_{w'} < \lambda \delta^{w'}(\epsilon) = \delta^w(\epsilon),$$

and consequently, since H_U has the FMP with respect to the weighting sequence w, we can conclude that $|H_U(\mathbf{z}) - H_U(\mathbf{s})| < \epsilon$. This shows that the implication

$$\|\mathbf{z} - \mathbf{s}\|_{w'} < \delta^{w'}(\epsilon) \Longrightarrow |H_U(\mathbf{z}) - H_U(\mathbf{s})| < \epsilon$$

holds, as required. ∎

6.4 Proof of Theorem 8

Since the elements in $\mathcal{R}$ have the FMP with respect to a given weighted norm $\|\cdot\|_w$, then so do those in $\mathcal{A}(\mathcal{R})$ since polynomial combinations of continuous elements of the form $H_{h_i}^{F_i} : (K_M, \|\cdot\|_w) \longrightarrow \mathbb{R}$ are also continuous. Therefore, under that hypothesis, $\mathcal{A}(\mathcal{R})$ is a polynomial subalgebra of the algebra $(C^0(K_M), \|\cdot\|_w)$ of real-valued continuous functions on $(K_M, \|\cdot\|_w)$. Since by hypothesis $\mathcal{A}(\mathcal{R})$ contains the constant functionals and separates the points in K_M and, by Lemma 2, the set $(K_M, \|\cdot\|_w)$ is compact, the Stone-Weierstrass theorem (Theorem 7.3.1 in Dieudonné (1969)) implies that $\mathcal{A}(\mathcal{R})$ is dense in $(C^0(K_M), \|\cdot\|_w)$, which concludes the proof. ■

6.5 Proof of Corollary 11

In order to show that the reservoir systems in $\mathcal{L}_\epsilon$ induce reservoir filters, we first show that they have the echo state property by using the following lemma, whose proof can be found in Grigoryeva and Ortega (2018).

Lemma 35 *Let $D_N \subset \mathbb{R}^N$ and $D_n \subset \mathbb{R}^n$ and let $F : D_N \times D_n \longrightarrow D_N$ be a continuous reservoir map. Suppose that F is a contraction map with contraction constant $0 < r < 1$, that is:*

$$\|F(\mathbf{x}, \mathbf{z}) - F(\mathbf{y}, \mathbf{z})\| \leq r \|\mathbf{x} - \mathbf{y}\|, \quad \textit{for all } \mathbf{x}, \mathbf{y} \in D_N \textit{ and all } \mathbf{z} \in D_n,$$

then the corresponding reservoir system has the echo state property.

We start now by noting that the condition $\sigma_{\max}(A) < 1 - \epsilon < 1$ implies that the reservoir map $F(\mathbf{x}, \mathbf{z}) := A\mathbf{x} + \mathbf{c}\mathbf{z}$ associated to (3.7) is a contracting map with constant $\sigma_{\max}(A)$ which, by hypothesis, is smaller than one. Indeed,

$$\|F(\mathbf{x}, \mathbf{z}) - F(\mathbf{y}, \mathbf{z})\| = \|A(\mathbf{x} - \mathbf{y})\| \leq \sigma_{\max}(A) \|\mathbf{x} - \mathbf{y}\| \quad \text{for all } \mathbf{x}, \mathbf{y} \in D_N \text{ and all } \mathbf{z} \in D_n.$$

By Lemma 35 we can conclude that this reservoir system has a reservoir filter associated that we now show is explicitly given by (3.9). We start by proving that the conditions $\sigma_{\max}(A) < 1 - \epsilon < 1$ and that the elements in K_M are uniformly bounded by a constant M imply that the infinite sum in (3.9) is convergent. Let $n, m \in \mathbb{N}$ be such that $n < m$ and let $S_n := \sum_{i=0}^{n} A^i \mathbf{c} z_{-i}$. Now:

$$\|S_n - S_m\| = \left\| \sum_{j=n+1}^{m} A^i \mathbf{c}\mathbf{z}_{-i} \right\| \leq \sum_{j=n+1}^{m} \|A\|_2^i \|\mathbf{c}\|_2 \|\mathbf{z}_{-i}\| \leq M \|\mathbf{c}\|_2 \sum_{j=n+1}^{m} \sigma_{\max}(A)^i$$
$$\leq M \|\mathbf{c}\|_2 \sum_{j=n+1}^{\infty} \sigma_{\max}(A)^i = M \|\mathbf{c}\|_2 \frac{\sigma_{\max}(A)^{n+1}}{1 - \sigma_{\max}(A)}.$$

The condition $\sigma_{\max}(A) < 1 - \epsilon < 1$ implies that $M \|\mathbf{c}\|_2 \frac{\sigma_{\max}(A)^{n+1}}{1 - \sigma_{\max}(A)} = M \frac{\sigma_{\max}(\mathbf{c})\sigma_{\max}(A)^{n+1}}{1 - \sigma_{\max}(A)} \to 0$ as $n \to \infty$ and hence $\{S_n\}_{n \in \mathbb{N}}$ is a Cauchy sequence in $\mathbb{R}^N$ that consequently converges.

The fact that the filter determined by the expression (3.9) is a solution of the recursions (3.7)-(3.8) is a straightforward verification. In order to carry it out, it suffices to use that the filter $U_h^{A,\mathbf{c}}(\mathbf{z})$ associated to the functional $H_h^{A,\mathbf{c}}(\mathbf{z})$ is given by

$$U_h^{A,\mathbf{c}}(\mathbf{z})_t = h \left(\sum_{i=0}^{\infty} A^i \mathbf{c}\mathbf{z}_{t-i} \right),$$

and that the time series $\widetilde{\mathbf{x}}_t$ defined by $\widetilde{\mathbf{x}}_t := \sum_{i=0}^{\infty} A^i \mathbf{c}\mathbf{z}_{t-i}$ satisfies the recursion relation (3.7).

We now verify by hand that the filters $U_h^{A,\mathbf{c}}$ are time-invariant. Let $\mathbf{z} \in K_M$ and $t, \tau \in \mathbb{N}$ arbitrary and let U_τ be the corresponding time delay operator, then:

$$\left(U_h^{A,\mathbf{c}} \circ U_\tau\right)(\mathbf{z})_t = \left(U_h^{A,\mathbf{c}}\left(U_\tau(\mathbf{z})\right)\right)_t = h\left(\sum_{i=0}^{\infty} A^i \mathbf{c} U_\tau(\mathbf{z})_{t-i}\right) = h\left(\sum_{i=0}^{\infty} A^i \mathbf{c} \mathbf{z}_{t-i-\tau}\right) \tag{6.4}$$

At the same time,

$$\left(U_\tau \circ U_h^{A,\mathbf{c}}\right)(\mathbf{z})_t = \left(U_\tau\left(U_h^{A,\mathbf{c}}(\mathbf{z})\right)\right)_t = U_h^{A,\mathbf{c}}(\mathbf{z})_{t-\tau} = h\left(\sum_{i=0}^{\infty} A^i \mathbf{c} \mathbf{z}_{t-\tau-i}\right),$$

which coincides with (6.4) and proves the time-invariance of $U_h^{A,\mathbf{c}}$.

The next step consists in showing that the elements in $\mathcal{L}_\epsilon$ are λ_ρ-exponential fading memory filters, with $\lambda_\rho := (1-\epsilon)^\rho$, for any $\rho \in (0,1)$, that is, $\mathcal{L}_\epsilon \subset \mathcal{R}_{w^\rho}$, with $w^\rho : \mathbb{N} \to (0,1]$ the sequence given by $w_t^\rho := (1-\epsilon)^{\rho t}$. Let $\|\cdot\|_{w^\rho}$ be the associated weighted norm in K_M and let $\mathbf{z} \in K_M$ be an arbitrary element. We start by noting that the continuity of the readout map $h : D_N \to \mathbb{R}$ implies that for any $\varepsilon > 0$ there exists an element $\delta(\varepsilon) > 0$ such that for any $\mathbf{v} \in D_N$ that satisfies

$$\left\|\mathbf{v} - \sum_{i=0}^{\infty} A^i \mathbf{c} \mathbf{z}_{t-i}\right\| < \delta(\varepsilon), \quad \text{then} \quad \left|h(\mathbf{v}) - h\left(\sum_{i=0}^{\infty} A^i \mathbf{c} \mathbf{z}_{t-i}\right)\right| < \varepsilon. \tag{6.5}$$

We now show that for any $\mathbf{s} \in K_M$ such that

$$\|\mathbf{s} - \mathbf{z}\|_{w^\rho} < \frac{\delta(\varepsilon)\left(1 - (1-\epsilon)^{1-\rho}\right)}{\sigma_{\max}(\mathbf{c})}, \quad \text{then} \quad \left|H_h^{A,\mathbf{c}}(\mathbf{s}) - H_h^{A,\mathbf{c}}(\mathbf{z})\right| < \varepsilon. \tag{6.6}$$

Indeed,

$$\begin{aligned}\left\|\sum_{i=0}^{\infty} A^i \mathbf{c} \mathbf{s}_{t-i} - \sum_{i=0}^{\infty} A^i \mathbf{c} \mathbf{z}_{t-i}\right\| &= \left\|\sum_{i=0}^{\infty} A^i \mathbf{c}(\mathbf{s}_{t-i} - \mathbf{z}_{t-i})\right\| \leq \sum_{i=0}^{\infty} \left\|A^i \mathbf{c}(\mathbf{s}_{t-i} - \mathbf{z}_{t-i})\right\| \\ &\leq \sum_{i=0}^{\infty} \sigma_{\max}(A^i) \left\|\mathbf{c}(\mathbf{s}_{t-i} - \mathbf{z}_{t-i})\right\| \leq \sum_{i=0}^{\infty} \sigma_{\max}(A)^i \left\|\mathbf{c}(\mathbf{s}_{t-i} - \mathbf{z}_{t-i})\right\| \leq \sum_{i=0}^{\infty} (1-\epsilon)^i \left\|\mathbf{c}(\mathbf{s}_{t-i} - \mathbf{z}_{t-i})\right\|.\end{aligned}$$

If we now use (2.5) in Lemma 1 and the hypothesis in (6.6), we can conclude that

$$\sum_{i=0}^{\infty} (1-\epsilon)^i \left\|\mathbf{c}(\mathbf{s}_{t-i} - \mathbf{z}_{t-i})\right\| \leq \sigma_{\max}(\mathbf{c}) \sum_{i=0}^{\infty} (1-\epsilon)^i \left\|(\mathbf{s}_{t-i} - \mathbf{z}_{t-i})\right\| \leq \frac{\sigma_{\max}(\mathbf{c}) \|\mathbf{s} - \mathbf{z}\|_{w^\rho}}{1 - (1-\epsilon)^{1-\rho}} < \delta(\varepsilon),$$

which proves the continuity of the map $H_h^{A,\mathbf{c}} : (K_M, \|\cdot\|_{w^\rho}) \longrightarrow \mathbb{R}$ and hence shows that $H_h^{A,\mathbf{c}}$ is a λ_ρ-exponential fading memory filter.

In order to establish the universality statement in the corollary we will proceed, as in the proof of Theorem 8, by showing that $\mathcal{L}_\epsilon$ is a polynomial algebra that contains the constant functionals and separates the points in K_M and then by invoking the Stone-Weierstrass theorem using the compactness of $(K_M, \|\cdot\|_{w^\rho})$.

In order to show that $(\mathcal{L}_\epsilon, \|\cdot\|_{w^\rho})$ is a polynomial algebra, notice first that if $A_1, A_2 \in \mathbb{M}_N$ are such that $\sigma_{\max}(A_1), \sigma_{\max}(A_2) < 1 - \epsilon$, then

$$\sigma_{\max}(A_1 \oplus A_2) = \max\left(\sigma_{\max}(A_1), \sigma_{\max}(A_2)\right) < 1 - \epsilon. \tag{6.7}$$

If we now take $\mathbf{c}_i \in \mathbb{M}_{N_i,n}$, $i \in \{1,2\}$ and h_1, h_2 two real-valued polynomials in N_1 and N_2 variables, respectively, we have by the first part of the corollary that we just proved that the filter functionals $H_{h_1}^{A_1,\mathbf{c}_1}$ and $H_{h_2}^{A_2,\mathbf{c}_2}$ are well defined. Additionally, by (3.3)-(3.4) so are the combinations $H_{h_1}^{A_1,\mathbf{c}_1} \cdot H_{h_2}^{A_2,\mathbf{c}_2}$ and $H_{h_1}^{A_1,\mathbf{c}_1} + \lambda H_{h_2}^{A_2,\mathbf{c}_2}$ that satisfy:

$$H_{h_1}^{A_1,\mathbf{c}_1} \cdot H_{h_2}^{A_2,\mathbf{c}_2} = H_{h_1 \cdot h_2}^{A_1 \oplus A_2, \mathbf{c}_1 \oplus \mathbf{c}_2}, \quad H_{h_1}^{A_1,\mathbf{c}_1} + \lambda H_{h_2}^{A_2,\mathbf{c}_2} = H_{h_1 \oplus \lambda h_2}^{A_1 \oplus A_2, \mathbf{c}_1 \oplus \mathbf{c}_2}, \quad \lambda \in \mathbb{R}. \tag{6.8}$$

Using the relations (6.8) and (6.7), we can conclude that both $H_{h_1}^{A_1,\mathbf{c}_1} \cdot H_{h_2}^{A_2,\mathbf{c}_2}$ and $H_{h_1}^{A_1,\mathbf{c}_1} + \lambda H_{h_2}^{A_2,\mathbf{c}_2}$ belong to $\mathcal{L}_\epsilon \subset \mathcal{R}_{w^\rho}$. This implies that $(\mathcal{L}_\epsilon, \|\cdot\|_{w^\rho})$ is a polynomial subalgebra of $(\mathcal{R}_{w^\rho}, \|\cdot\|_{w^\rho})$

Since $\mathcal{L}_\epsilon$ contains the constant functionals (just take constant readout maps h), in order to conclude the proof, it is enough to show that the elements in $\mathcal{L}_\epsilon$ separate points in K_M. In the proof of this statement we need the following elementary fact about analytic functions.

Lemma 36 *Let $M > 0$ and let $\mathbf{z} \in [-M, M]^{\mathbb{Z}_-}$. Define the real valued function $f_{\mathbf{z}}(x) := \sum_{j=0}^{\infty} z_{-j} x^j$. This function is real analytic in the interval $(-1,1)$. Moreover, if $\mathbf{z} \neq \mathbf{0}$, then there exists a point $x_0 \in (-1,1)$ such that $f_{\mathbf{z}}(x_0) \neq 0$.*

Proof of the lemma. We note first that for any $x \in (-1,1)$ and any $s \in \mathbb{N}$ we have that

$$\left| \sum_{j=0}^{s} z_{-j} x^j \right| \leq \sum_{j=0}^{s} |z_{-j}| \left| x^j \right| \leq M \sum_{j=0}^{s} |x|^j \leq \frac{M}{1-|x|}.$$

Taking the limit $s \to \infty$, we obtain that

$$|f_{\mathbf{z}}(x)| \leq \frac{M}{1-|x|}, \quad \text{for all } x \in (-1,1),$$

which proves the first claim in the lemma. Now, by the uniqueness theorem for the representation of analytic functions by power series (see Brown and Churchill (2009), page 217), the series $\sum_{j=0}^{\infty} z_{-j} x^j$ is the Taylor expansion around 0 of $f_{\mathbf{z}}(x)$. Since $\mathbf{z} \neq \mathbf{0}$ by hypothesis, some of the derivatives of $f_{\mathbf{z}}(x)$ are non-zero and hence this function cannot be flat, which implies that there exists a point $x_0 \in (-1,1)$ such that $f_{\mathbf{z}}(x_0) \neq 0$. ▼

We now show that the elements in $\mathcal{L}_\epsilon$ separate points in K_M. Take $\mathbf{z}_1, \mathbf{z}_2 \in K_M \subset (\mathbb{R}^n)^{\mathbb{Z}_-}$ such that $\mathbf{z}_1 \neq \mathbf{z}_2$ and let $A \in \mathbb{M}(n,n)$, with $\sigma_{\max}(A) < 1 - \epsilon$, and $\mathbf{c} := \mathbb{I}_n$. Let $U^{A,\mathbf{c}} : K_M \longrightarrow (\mathbb{R}^n)^{\mathbb{Z}_-}$ be the linear filter associated to A and $\mathbf{c}$ via the recursion (3.7). Using the preceding arguments we have that

$$U^{A,\mathbf{c}}(\mathbf{z})_t = \sum_{j=0}^{\infty} A^j \mathbf{z}_{t-j}. \tag{6.9}$$

Since $\mathbf{z}_1 \neq \mathbf{z}_2$, then there exists and index $i \in \{1, \ldots, n\}$ and $t \in \mathbb{Z}_-$ such that $\left(z_1^i\right)_t \neq \left(z_2^i\right)_t$. Let now $b \in (-1+\epsilon, 1-\epsilon)$ and let $A_b := \operatorname{diag}(0, \ldots, 0, b, 0, \ldots, 0) \in \mathbb{D}_n$ be the matrix that has the element b in the i-th entry. It is easy to see using (6.9) that

$$U^{A_b,\mathbf{c}}(\mathbf{z})_t = \left(0, \ldots, 0, \sum_{j=0}^{\infty} b^j z_{t-j}^i, 0, \ldots, 0\right)^\top, \quad \text{with} \quad \sum_{j=0}^{\infty} b^j z_{t-j}^i \quad \text{in the } i\text{-th entry.} \tag{6.10}$$

Let $\mathbf{s} := \mathbf{z}_1 - \mathbf{z}_2 \neq \mathbf{0}$. Notice that by (6.10) we have that $U^{A_b,\mathbf{c}}(\mathbf{s})_0 = \left(0, \ldots, 0, \sum_{j=0}^{\infty} b^j s_{-j}^i, 0, \ldots, 0\right)^\top$. Given that the vector $\mathbf{s}^i \in \mathbb{R}^{\mathbb{Z}_-}$ is non-zero, Lemma 36, implies the existence of an element $b_0 \in$

$(-1+\epsilon, 1-\epsilon)$ such that $U^{A_{b_0},\mathbf{c}}(\mathbf{s})_0 \neq \mathbf{0}$, which is equivalent to $U^{A_{b_0},\mathbf{c}}(\mathbf{z}_1)_0 \neq U^{A_{b_0},\mathbf{c}}(\mathbf{z}_2)_0$. Using the polynomial $h(\mathbf{x}) := x_i \in \mathbb{R}$, the previous relation implies that $U_h^{A_{b_0},\mathbf{c}}(\mathbf{z}_1)_0 \neq U_h^{A_{b_0},\mathbf{c}}(\mathbf{z}_2)_0$ or, equivalently,

$$H_h^{A_{b_0},\mathbf{c}}(\mathbf{z}_1) \neq H_h^{A_{b_0},\mathbf{c}}(\mathbf{z}_2), \quad \text{as required.}$$

We conclude the proof by establishing the universality the families $\mathcal{DL}_\epsilon$ and $\mathcal{NL}$ formed by the linear reservoir filters generated by diagonal and nilpotent matrices, respectively. First, in the case of $\mathcal{DL}_\epsilon$, the statement is a consequence of (6.8) and of the fact that when the matrices A_1 and A_2 are diagonal, then the matrix associated to the linear map $A_1 \oplus A_2$ is also diagonal. Additionally, notice that the point separation property for $\mathcal{L}_\epsilon$ has been proved using diagonal matrices in (6.10) and hence it also holds for $\mathcal{DL}_\epsilon$. The claim follows from the Stone-Weierstrass theorem.

Finally, in the case of $\mathcal{NL}$, the proof also follows from (6.8) since it is straightforward to see that when the matrices A_1 and A_2 are nilpotent, then the matrix associated to the linear map $A_1 \oplus A_2$ is also nilpotent. It is only the point separation property of $\mathcal{N}$ that requires a separate argument that we provide in the following lines. Let $\mathbf{z}_1, \mathbf{z}_2 \in K_M$ such that $\mathbf{z}_1 \neq \mathbf{z}_2$ and let $t_0 \in \mathbb{N}$ be the first time index for which $(\mathbf{z}_1)_{-t_0} \neq (\mathbf{z}_2)_{-t_0}$, that is, $(\mathbf{z}_1)_{-t} = (\mathbf{z}_2)_{-t}$, for all $t \in \{0, 1, \ldots, t_0-1\}$. Let now $i_0 \in \{1, \ldots, n\}$ be such that $\left(z_1^{i_0}\right)_{-t_0} \neq \left(z_2^{i_0}\right)_{-t_0}$. Let now $A_{t_0+1} \in \mathrm{Nil}_{t_0+1}^{t_0+1}$ be the upper shift matrix in dimension t_0+1, that is, $A_{t_0+1} \in \mathbb{M}_{t_0+1}$ is by definition a superdiagonal matrix with a diagonal of ones above the main diagonal, and construct an element $\mathbf{c} \in \mathbb{M}_{t_0+1,n}$ whose last row is given by a vector of zeros with the exception of a one in the entry i_0. The nilpotency of A_{t_0+1} implies

$$U^{A_{t_0+1},\mathbf{c}}(\mathbf{z})_0 = \sum_{j=0}^{t_0} A_{t_0+1}^j \mathbf{c}\mathbf{z}_{-j}.$$

When we apply this expression to $\mathbf{z}_1$ and $\mathbf{z}_2$, since $(\mathbf{z}_1)_{-t} = (\mathbf{z}_2)_{-t}$, for all $t \in \{0, 1, \ldots, t_0-1\}$, we obtain that

$$U^{A_{t_0+1},\mathbf{c}}(\mathbf{z}_1 - \mathbf{z}_2)_0 = A_{t_0+1}^{t_0}\mathbf{c}(\mathbf{z_1} - \mathbf{z_2})_{-t_0} = \left(0, \ldots, 0, \left(z_1^{i_0}\right)_{-t_0} - \left(z_2^{i_0}\right)_{-t_0}\right)^\top \neq \mathbf{0}.$$

Using the polynomial $h(\mathbf{x}) := x_{t_0+1}$, this relation implies that $U_h^{A_{t_0+1},\mathbf{c}}(\mathbf{z}_1)_0 \neq U_h^{A_{t_0+1},\mathbf{c}}(\mathbf{z}_2)_0$ or, equivalently, $H_h^{A_{t_0+1},\mathbf{c}}(\mathbf{z}_1) \neq H_h^{A_{t_0+1},\mathbf{c}}(\mathbf{z}_2)$, as required. ■

6.6 Proof of Proposition 14

We start by noting, as we did in the proof of Corollary 11, that the condition (3.13) implies that the reservoir map associated to (3.11) is a contraction and hence, by Lemma 35, it satisfies the echo state property and has a well-defined associated filter.

We now prove that the condition (3.13) implies the convergence of the series in the expression (3.14). Let $K_1 := \max_{z \in I} \|p(z)\|_2 = \max_{z \in I} \sigma_{\max}(p(z)) < 1$ and $K_2 := \max_{z \in I} \|q(z)\|_2 = \max_{z \in I} \sigma_{\max}(q(z))$; notice that K_1 and K_2 are well-defined due to the compactness of I. Let now $n, m \in \mathbb{N}$ be such that $n < m$ and let $S_n := \sum_{j=0}^{n} \left(\prod_{k=0}^{j-1} p(z_{t-k})\right) q(z_{t-j}) \in \mathbb{R}^N$. Then,

$$\begin{aligned} \|S_n - S_m\| &= \left\| \sum_{j=n+1}^{m} \left(\prod_{k=0}^{j-1} p(z_{t-k})\right) q(z_{t-j}) \right\| \leq \sum_{j=n+1}^{m} \left\| \prod_{k=0}^{j-1} p(z_{t-k}) \right\|_2 \|q(z_{t-j})\| \\ &\leq \sum_{j=n+1}^{m} \prod_{k=0}^{j-1} \|p(z_{t-k})\|_2 \|q(z_{t-j})\| \leq K_2 \sum_{j=n+1}^{m} K_1^j \leq K_2 \sum_{j=n+1}^{\infty} K_1^j = \frac{K_2 K_1^{n+1}}{1-K_1}. \end{aligned}$$

The condition $K_1 < 1$ implies that $\frac{K_2 K_1^{n+1}}{1-K_1} \to 0$ as $n \to \infty$ and hence $\{S_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{R}^N$ that consequently converges. This proves the convergence of the infinite series in (3.14) and the causal character of the filter that it defines. The time-invariance can also be easily established by mimicking the verification that we carried out in the proof of Corollary 11. We now prove that (3.14) is indeed a solution of (3.11):

$$p(z_t)\mathbf{x}_{t-1} + q(z_t) = p(z_t)\left(\sum_{j=0}^{\infty}\left(\prod_{k=0}^{j-1} p(z_{t-1-k})\right) q(z_{t-1-j})\right) + q(z_t) = q(z_t) + p(z_t)q(z_{t-1})$$
$$+ p(z_t)p(z_{t-1})q(z_{t-2}) + p(z_t)p(z_{t-1})p(z_{t-2})q(z_{t-3}) + \cdots = \sum_{j=0}^{\infty}\left(\prod_{k=0}^{j-1} p(z_{t-k})\right) q(z_{t-j}) = \mathbf{x}_t.$$

We conclude by proving the inequality in (3.16). Note first that for any $m \in \mathbb{N}$,

$$\left\|\sum_{j=0}^{m}\left(\prod_{k=0}^{j-1} p(z_{t-k})\right) q(z_{t-j})\right\| \leq \sum_{j=0}^{m}\left\|\prod_{k=0}^{j-1} p(z_{t-k})\right\|_2 \|q(z_{t-j})\|$$
$$\leq \sum_{j=0}^{m}\prod_{k=0}^{j-1} \|p(z_{t-k})\|_2 \|q(z_{t-j})\| \leq \frac{K_2\left(1-K_1^{m+1}\right)}{1-K_1},$$

and hence, by the continuity of the norm and for any $t \in \mathbb{Z}$:

$$\|\mathbf{x}_t\| = \lim_{m\to\infty}\left\|\sum_{j=0}^{m}\left(\prod_{k=0}^{j-1} p(z_{t-k})\right) q(z_{t-j})\right\| \leq \lim_{m\to\infty}\frac{K_2\left(1-K_1^{m+1}\right)}{1-K_1} = \frac{K_2}{1-K_1}. \quad \blacksquare$$

6.7 Proof of Lemma 15

(i) $\Longrightarrow$ (ii): $\|A_0\|_2 + \|A_1\|_2 + \cdots + \|A_{n_1}\|_2 < \sum_{i=0}^{n_1} \lambda = \lambda(n_1+1) < 1$.
(ii) $\Longrightarrow$ (iii): $\|p(z)\|_2 = \|A_0 + zA_1 + z^2A_2 + \cdots + z^{n_1}A_{n_1}\|_2 \leq \|A_0\|_2 + |z|\|A_1\|_2 + |z^2|\|A_2\|_2 + \cdots + |z^{n_1}|\|A_{n_1}\|_2 < \|A_0\|_2 + \|A_1\|_2 + \cdots + \|A_{n_1}\|_2 < 1$. $\blacksquare$

6.8 Proof of Proposition 16

We start by formulating and proving an elementary result that will be needed later on.

Lemma 37 *Let $\mathbf{f} : U \subset \mathbb{R}^n \longrightarrow \mathbb{M}_m$ be a differentiable function defined on the convex set U. For any $\mathbf{z} \in U$ denote by $\partial_i \mathbf{f}(\mathbf{z}) \in \mathbb{M}_m$ the matrix containing the partial derivatives of the components of $\mathbf{f}$ with respect to their ith-entry, $i \in \{1, \ldots, n\}$. Then, for any $\mathbf{x}, \mathbf{y} \in U$ we have:*

$$\|\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x})\|_2 \leq \sqrt{nm} \max_{i\in\{1,\ldots,n\}}\left(\sup_{\mathbf{z}\in U}\{\|\partial_i \mathbf{f}(\mathbf{z})\|_2\}\right)\|\mathbf{x}-\mathbf{y}\|. \tag{6.11}$$

Proof. Given $A = (A_{i,j}) \in \mathbb{M}_{n,m}$, let $\|A\|_F := \operatorname{tr}\left(A^\top A\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} A_{i,j}^2$ be its Frobenius norm. Recall (see Theorem 5.6.34 and Exercise 5.6.P24 in Horn and Johnson (2013)) that

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{r}\,\|A\|_2, \tag{6.12}$$

where r is the rank of A. Consider now $\mathbf{x}, \mathbf{y} \in U$ arbitrary and let $D\mathbf{f}(\mathbf{z}) : \mathbb{R}^n \longrightarrow \mathbb{M}_m$ be the differential of $\mathbf{f}$ evaluated at $\mathbf{z} \in U$. The convexity of U implies that the Mean Value Inequality holds (see Theorem 2.4.8 in Abraham et al. (1988)) and hence:

$$\|\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x})\|_F \leq \sup_{t\in[0,1]} \{\|D\mathbf{f}((1-t)\mathbf{x} + t\mathbf{y})\|_2 \|\mathbf{x} - \mathbf{y}\|\}. \tag{6.13}$$

The first inequality in (6.12) and (6.13) imply that

$$\|\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x})\|_2 \leq \sup_{\mathbf{z}\in U}\{\|D\mathbf{f}(\mathbf{z})\|_2 \|\mathbf{x} - \mathbf{y}\|\}. \tag{6.14}$$

At the same time, notice that by (6.12)

$$\|D\mathbf{f}(\mathbf{z})\|_2^2 \leq \|D\mathbf{f}(\mathbf{z})\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m} \partial_i f_{jk}^2(\mathbf{z}) = \sum_{i=1}^{n} \|\partial_i \mathbf{f}(\mathbf{z})\|_F^2$$
$$\leq m \sum_{i=1}^{n} \|\partial_i \mathbf{f}(\mathbf{z})\|_2^2 \leq mn \max_{i\in\{1,\ldots,n\}} \left(\|\partial_i \mathbf{f}(\mathbf{z})\|_2^2\right).$$

This inequality, together with (6.14), imply the statement (6.11) since the maximum and the supremum can be trivially exchanged. ▼

We now carry out the proof of the proposition under the hypothesis **(iii)** in Lemma 15 which is implied by the other two. The modifications necessary to establish the result under the other two hypotheses are straightforward. Consider two arbitrary elements $\mathbf{z}, \mathbf{s} \in I^{\mathbb{Z}_-}$. Then, by the Cauchy-Schwarz and Minkowski inequalities:

$$\left|H^{p,q}_{\mathbf{W}}(\mathbf{z}) - H^{p,q}_{\mathbf{W}}(\mathbf{s})\right| = \left|\mathbf{W}^\top \left[\sum_{j=0}^{\infty}\left(\left(\prod_{k=0}^{j-1} p(z_{-k})\right) q(z_{-j}) - \left(\prod_{k=0}^{j-1} p(s_{-k})\right) q(s_{-j})\right)\right]\right|$$
$$\leq \|\mathbf{W}\| \sum_{j=0}^{\infty} \left\|a_j(\underline{z_{-j+1}})q(z_{-j}) - a_j(\underline{s_{-j+1}})q(s_{-j})\right\|, \quad \text{where} \quad a_j(\underline{z_{-j+1}}) := \prod_{k=0}^{j-1} p(z_{-k}). \tag{6.15}$$

We now bound the right hand side of (6.15) as follows:

$$\sum_{j=0}^{\infty} \left\|a_j(\underline{z_{-j+1}})q(z_{-j}) - a_j(\underline{s_{-j+1}})q(s_{-j})\right\|$$
$$= \sum_{j=0}^{\infty} \left\|a_j(\underline{z_{-j+1}})q(z_{-j}) + a_j(\underline{z_{-j+1}})q(s_{-j}) - a_j(\underline{z_{-j+1}})q(s_{-j}) - a_j(\underline{s_{-j+1}})q(s_{-j})\right\|$$
$$\leq \sum_{j=0}^{\infty} \left\|a_j(\underline{z_{-j+1}})\right\|_2 \|q(z_{-j}) - q(s_{-j})\| + \left\|a_j(\underline{z_{-j+1}}) - a_j(\underline{s_{-j+1}})\right\|_2 \|q(s_{-j})\| \tag{6.16}$$

If L_q is a Lipschitz constant of $q : I \longrightarrow \mathbb{R}^N$ then

$$\left\|a_j(\underline{z_{-j+1}})\right\|_2 \|q(z_{-j}) - q(s_{-j})\| \leq M_p^j L_q \left|z_{-j} - s_{-j}\right|, \tag{6.17}$$

which inserted in (6.16) and in (6.15) implies that

$$\left|H^{p,q}_{\mathbf{W}}(\mathbf{z}) - H^{p,q}_{\mathbf{W}}(\mathbf{s})\right| \leq \|\mathbf{W}\| L_q \left[\sum_{j=0}^{\infty} M_p^j \left|z_{-j} - s_{-j}\right| + \sum_{j=0}^{\infty} \left\|a_j(\underline{z_{-j+1}}) - a_j(\underline{s_{-j+1}})\right\|_2\right] \tag{6.18}$$

We now bound above the second summand in (6.18) using the inequality (6.11) in the statement of Lemma 37 as well as the following identity:

$$a_j(\underline{z_{-j+1}}) - a_j(\underline{s_{-j+1}}) = \sum_{l=0}^{j-1}(p(s_0)\cdots p(s_{-(l-1)})\cdot p(z_{-l})\cdot p(z_{-(l+1)})\cdots p(z_{-(j-1)}) \\ - p(s_0)\cdots p(s_{-(l-1)})\cdot p(s_{-l})\cdot p(z_{-(l+1)})\cdots p(z_{-(j-1)})). \quad (6.19)$$

This equality simply follows from writing:

$$\begin{aligned} a_j(\underline{z_{-j+1}})-a_j(\underline{s_{-j+1}}) &= \prod_{l=0}^{j-1} p(z_{-l})-\prod_{l=0}^{j-1} p(s_{-l}) = p(z_0)p(z_{-1})\cdots p(z_{-(j-1)})-p(s_0)p(s_{-1})\cdots p(s_{-(j-1)}) \\ &= p(z_0)p(z_{-1})\cdots p(z_{-(j-1)}) - p(s_0)p(s_{-1})\cdots p(s_{-(j-1)}) \\ &+ \Big\{ p(s_0)p(z_{-1})\cdots p(z_{-(j-1)}) - p(s_0)p(z_{-1})\cdots p(z_{-(j-1)}) \\ &+ p(s_0)p(s_{-1})p(z_{-2})\cdots p(z_{-(j-1)}) - p(s_0)p(s_{-1})p(z_{-2})\cdots p(z_{-(j-1)}) \\ +\cdots+ & p(s_0)\cdots p(s_{-(l-1)})p(z_{-l})p(z_{-(l+1)})\cdots p(z_{-(j-1)})-p(s_0)\cdots p(s_{-(l-1)})p(z_{-l})p(z_{-(l+1)})\cdots p(z_{-(j-1)}) \\ &+\cdots+ p(s_0)\cdots p(s_{-(j-2)})p(z_{-(j-1)}) - p(s_0)\cdots p(s_{-(j-2)})p(z_{-(j-1)}) \Big\} \\ &= \sum_{l=0}^{j-1}(p(s_0)\cdots p(s_{-(l-1)})\cdot p(z_{-l})\cdot p(z_{-(l+1)})\cdots p(z_{-(j-1)}) \\ &\qquad - p(s_0)\cdots p(s_{-(l-1)})\cdot p(s_{-l})\cdot p(z_{-(l+1)})\cdots p(z_{-(j-1)})), \end{aligned}$$

where the $2(j-1)$ summands inside the braces are obtained by adding and subtracting polynomials recursively constructed out of $a_j(\underline{z_{-j+1}})$ by changing the variables of the first k factors, $k \in \{1, \cdots, j-1\}$. We then combine all the $(2k-1)$-th with the $(2k+2)-th$ summands of the resulting expression in order to obtain the first $j-1$ terms in the sum in (6.19). Then the last j-th term results from combining the second with the one before last summands, that is, $p(s_0)p(s_{-1})\cdots p(s_{-(j-1)})$ and $p(s_0)\cdots p(s_{-(j-2)})p(z_{-(j-1)})$, respectively.

Using the relation (6.19) we can write:

$$\begin{aligned} \left\|a_j(\underline{z_{-j+1}}) - a_j(\underline{s_{-j+1}})\right\|_2 &\leq \sum_{l=0}^{j-1} \left\|p(s_0)\cdots p(s_{-(l-1)})\cdot (p(z_{-l})-p(s_{-l}))\cdot p(z_{-(l+1)})\cdots p(z_{-(j-1)})\right\|_2 \\ &\leq \sum_{l=0}^{j-1} \|p(s_0)\|_2\cdots \left\|p(s_{-(l-1)})\right\|_2 \cdot \|p(z_{-l})-p(s_{-l})\|_2 \cdot \left\|p(z_{-(l+1)})\right\|_2 \cdots \left\|p(z_{-(j-1)})\right\|_2 \\ &\leq M_p^{j-1}\sqrt{N}\sup_{z\in I}\{\|p'(z)\|_2\} \sum_{l=1}^{j} |z_{-j+l}-s_{-j+l}|, \end{aligned}$$

where the last inequality is a consequence of (6.11). Let $M_{p'} := \sqrt{N}\sup_{z\in I}\{\|p'(z)\|_2\}$, then

$$\left\|a_j(\underline{z_{-j+1}}) - a_j(\underline{s_{-j+1}})\right\|_2 \leq \frac{M_{p'}}{M_p} M_p^j \sum_{l=1}^{j} |z_{-j+l}-s_{-j+l}| = \frac{M_{p'}}{M_p} \sum_{l=1}^{j} M_p^l M_p^{j-l} \left|z_{-(j-l)}-s_{-(j-l)}\right|$$

Since the last term in this inequality is one summand of the Cauchy product of the series with general terms M_p^j and $M_p^j \left|z_{-j} - s_{-j}\right|$ and these two series are absolutely convergent (recall the statement (2.4)), we can conclude (see, for instance, §8.24 in Apostol (1974)) that

$$\sum_{j=0}^{\infty} \left\| a_j(\underline{z_{-j+1}}) - a_j(\underline{s_{-j+1}}) \right\|_2 \leq \frac{M_{p'}}{M_p} \sum_{j=0}^{\infty} \sum_{l=1}^{j} M_p^l M_p^{j-l} \left| z_{-(j-l)} - s_{-(j-l)} \right|$$
$$= \frac{M_{p'}}{M_p} \frac{1}{1-M_p} \sum_{j=0}^{\infty} M_p^j \left| z_{-j} - s_{-j} \right|.$$

If we now substitute this relation in (6.18) and we use Lemma 1 with weighting sequences $w_t^\rho := M_p^{\rho t}$, for any $\rho \in (0,1)$, we obtain that:

$$\begin{aligned} \left|H_{\mathbf{W}}^{p,q}(\mathbf{z}) - H_{\mathbf{W}}^{p,q}(\mathbf{s})\right| &\leq \|\mathbf{W}\| L_q \left(1 + \frac{M_{p'}}{M_p} \frac{1}{1-M_p}\right) \sum_{j=0}^{\infty} M_p^j \left|z_{-j} - s_{-j}\right| \\ &\leq \|\mathbf{W}\| L_q \left(1 + \frac{M_{p'}}{M_p} \frac{1}{1-M_p}\right) \left(\frac{1}{1-M_p^{1-\rho}}\right) \|\mathbf{z} - \mathbf{s}\|_{w^\rho}, \end{aligned}$$

which proves the continuity of the map $H_{\mathbf{W}}^{p,q} : (I^{\mathbb{Z}_-}, \| \cdot \|_{w^\rho}) \longrightarrow \mathbb{R}$, as required. ■

6.9 Proof of Proposition 17

We first recall that since by hypothesis the reservoir functionals $H_{\mathbf{W}_1}^{p_1,q_1}, H_{\mathbf{W}_2}^{p_2,q_2}$ are well-defined then, by the comments that follow (3.5), so are $H_{\mathbf{W}_1}^{p_1,q_1} + \lambda H_{\mathbf{W}_2}^{p_2,q_2}$ and $H_{\mathbf{W}_1}^{p_1,q_1} \cdot H_{\mathbf{W}_2}^{p_2,q_2}$.

The proof of (**i**) is a straightforward verification. As to (**ii**), denote first by y_t^1, y_t^2 and $\mathbf{x}_t^1, \mathbf{x}_t^2$ the outputs and the state variables, respectively, of the SAS corresponding to the two functionals that we are considering. We note first that by (3.12):

$$y_t^1 \cdot y_t^2 = \mathbf{W}_1^\top \mathbf{x}_t^1 \cdot \mathbf{W}_2^\top \mathbf{x}_t^2 = \left(\mathbf{W}_1 \otimes \mathbf{W}_2\right)^\top (\mathbf{x}_t^1 \otimes \mathbf{x}_t^2).$$

Using (3.11) it can be readily verified that the time evolution of the tensor product $\mathbf{x}_t^1 \otimes \mathbf{x}_t^2$ is given by

$$\begin{aligned} \mathbf{x}_t^1 \otimes \mathbf{x}_t^2 &= (p_1(z_t) \otimes p_2(z_t))(\mathbf{x}_{t-1}^1 \otimes \mathbf{x}_{t-1}^2) + p_1(z_t)\mathbf{x}_{t-1}^1 \otimes q_2(z_t) + q_1(z_t) \otimes p_2(z_t)\mathbf{x}_{t-1}^2 + q_1(z_t) \otimes q_2(z_t), \\ &= (p_1 \otimes p_2)(z_t)(\mathbf{x}_{t-1}^1 \otimes \mathbf{x}_{t-1}^2) + p_1(z_t)\mathbf{x}_{t-1}^1 \otimes q_2(z_t) + q_1(z_t) \otimes p_2(z_t)\mathbf{x}_{t-1}^2 + (q_1 \otimes q_2)(z_t), \end{aligned}$$

which proves (3.23) and hence (3.22).

In order to show that the reservoir functionals on the right hand side of (3.21) and (3.22) are well-defined we prove the following lemma.

Lemma 38 *Let $p_1(z) \in \mathbb{M}_{N_1,M_1}[z]$ and $p_2(z) \in \mathbb{M}_{N_2,M_2}[z]$ be two polynomials with matrix coefficients and assume that they satisfy that $\|p_1(z)\|_2 < 1-\epsilon$ and $\|p_2(z)\|_2 < 1-\epsilon$ for all $z \in I := [-1,1]$ and a given $0 < \epsilon > 1$. Then:*

(i) $\|p_1 \oplus p_2(z)\|_2 < 1-\epsilon$,

(ii) $\|p_1 \otimes p_2(z)\|_2 < 1-\epsilon$,

for all $z \in I := [-1,1]$.

Proof of the lemma. Let $\mathbf{x} = \mathbf{x}_1 \oplus \mathbf{x}_2 \in \mathbb{R}^{M_1} \oplus \mathbb{R}^{M_2}$. Then, in order to prove part **(i)** note that

$$\begin{aligned}\|(p_1 \oplus p_2)(z) \cdot \mathbf{x}\|^2 &= \|(p_1(z) \cdot \mathbf{x}_1, p_2(z) \cdot \mathbf{x}_2)\|^2 = \|p_1(z) \cdot \mathbf{x}_1\|^2 + \|p_2(z) \cdot \mathbf{x}_2\|^2 \\ &\leq \|p_1(z)\|_2^2\|\mathbf{x}_1\|^2 + \|p_2(z)\|_2^2\|\mathbf{x}_2\|^2 \leq (1-\epsilon)^2\left(\|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|^2\right) = (1-\epsilon)^2\|\mathbf{x}\|^2.\end{aligned}$$

This inequality implies that

$$\|p_1 \oplus p_2(z)\|_2 = \sup_{\mathbf{x}\neq\mathbf{0}}\left\{\frac{\|(p_1 \oplus p_2)(z) \cdot \mathbf{x}\|}{\|\mathbf{x}\|}\right\} \leq \sup_{\mathbf{x}\neq\mathbf{0}}\left\{\frac{(1-\epsilon)\|\mathbf{x}\|}{\|\mathbf{x}\|}\right\} = 1-\epsilon, \quad \text{as required.}$$

As to the statement in part **(ii)**:

$$\|p_1 \otimes p_2(z)\|_2 = \sigma_{\max}(p_1 \otimes p_2(z)) = \sigma_{\max}(p_1(z))\sigma_{\max}(p_2(z)) = \|p_1(z)\|_2\|p_2(z)\|_2 < (1-\epsilon)^2 < (1-\epsilon). \blacktriangledown$$

Now, the first part of this lemma and Proposition 14 guarantee that $H^{p_1\oplus p_2, q_1\oplus q_2}_{\mathbf{W}_1\oplus\lambda\mathbf{W}_2}$ is well-defined. The same conclusion holds for $H^{p,q_1\oplus q_2\oplus(q_1\otimes q_2)}_{\mathbf{0}\oplus\mathbf{0}\oplus(\mathbf{W}_1\otimes\mathbf{W}_2)}$ because due to the block diagonal character of (3.23) then $\sigma_{\max}(p(z)) = \sigma_{\max}((p_1(z) \oplus p_2(z) \oplus (p_1 \otimes p_2)(z)) = \|p_1(z) \oplus p_2(z) \oplus (p_1 \otimes p_2)(z)\|_2$. By parts **(i)** and **(ii)** in Lemma 38 we can conclude that $\|p(z)\|_2 < 1-\epsilon$ for all $z \in [-1, 1]$ and, again by Proposition 14, the reservoir functional $H^{p,q_1\oplus q_2\oplus(q_1\otimes q_2)}_{\mathbf{0}\oplus\mathbf{0}\oplus(\mathbf{W}_1\otimes\mathbf{W}_2)}$ is well-defined. ■

6.10 Proof of Theorem 19

Note first that the hypothesis $M_p < 1-\epsilon < 1$ on the polynomials p associated to the elements in $\mathcal{S}_\epsilon$ implies, by Propositions 14 and 16, that this family is made of time-invariant reservoir filters that have the FMP with respect to weighting sequences of the form $w^p_t := M_p^{\rho t}$, $\rho \in (0, 1)$. Additionally, using Lemma 7 and the hypothesis $M_p < 1-\epsilon$, for a fixed given $\epsilon \in (0, 1)$, we can conclude that all the reservoir filters in $\mathcal{S}_\epsilon$ have the FMP with the common weighting sequence $w^\rho_t := (1-\epsilon)^{\rho t}$, $\rho \in (0, 1)$.

The elements in $\mathcal{S}_\epsilon$ form a polynomial algebra as a consequence of Lemma 38 and Proposition 17. Moreover, the family $\mathcal{S}_\epsilon$ has the point separation property and contains all the constant functionals. Indeed, since $\mathcal{S}_\epsilon$ includes the linear family $\mathcal{L}_\epsilon$, we recall that in Appendix 6.5 we proved that given $\mathbf{z}_1, \mathbf{z}_2 \in K_M \subset (\mathbb{R}^n)^{\mathbb{Z}_-}$ such that $\mathbf{z}_1 \neq \mathbf{z}_2$, there exists $A \in \mathbb{M}(n, n)$, with $\sigma_{\max}(A) < 1-\epsilon$ and $\mathbf{c} := \mathbb{I}_n$ such that $U^{A,\mathbf{c}}(\mathbf{z}_1)_0 \neq U^{A,\mathbf{c}}(\mathbf{z}_2)_0$. The point separation property follows from choosing any vector $\mathbf{W} \in \mathbb{R}^N$ such that $\mathbf{W}^\top(U^{A,\mathbf{c}}(\mathbf{z}_1))_0 \neq \mathbf{W}^\top(U^{A,\mathbf{c}}(\mathbf{z}_2))_0$, which implies that $U^{A,\mathbf{c}}_{\mathbf{W}}(\mathbf{z}_1)_0 \neq U^{A,\mathbf{c}}_{\mathbf{W}}(\mathbf{z}_2)_0$ and hence $H_{U^{A,\mathbf{c}}_{\mathbf{W}}}(\mathbf{z}_1) \neq H_{U^{A,\mathbf{c}}_{\mathbf{W}}}(\mathbf{z}_2)$, as required.

All the constant functionals can be obtained by taking for p the zero polynomial and for q the constant polynomials (q has degree zero). In that case, the state variables are a constant sequence $\mathbf{x}_t = q$ and the associated functional is the constant map $H^{0,q}_{\mathbf{W}}(\mathbf{z}) = \mathbf{W}^\top q$, for all $\mathbf{z} \in K_M$.

The universality result follows hence from the Stone-Weierstrass Theorem and the compactness of $(I^{\mathbb{Z}_-}, \|\cdot\|_{w^\rho})$ established in Lemma 2.

Finally, we prove the statement regarding the family $\mathcal{NS}_\epsilon$ determined by nilpotent polynomials p. First, by expressions (3.21), (3.22), and (3.23), it is easy to show that this family is a polynomial algebra. The only point that requires some detail is the fact that the k-th power of the polynomial p in (3.23) that is obtained in the product of the two SAS reservoir functionals $H^{p_1,q_1}_{\mathbf{W}_1}$ and $H^{p_2,q_2}_{\mathbf{W}_2}$ is given by

$$p^k(z) := \begin{pmatrix} p_1^k(z) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & p_2^k(z) & \mathbf{0} \\ p_1^k \otimes q_2^{k-1}(z) & q_1^{k-1} \otimes p_2^k(z) & p_1^k \otimes p_2^k(z) \end{pmatrix},$$

which shows that if p_1 and p_2 are nilpotent then so is the associated polynomial p. The point separation property is, again, inherited from the proof of linear case provided in the Appendix 6.5. ■

6.11 Proof of Lemma 24

(i) Let $A := \left\{\rho \in \overline{\mathbb{R}_+} \mid \|\mathbf{X}\|_B \leq \rho \quad \text{almost surely}\right\}$. It suffices to show that $\|\mathbf{X}\|_{L^\infty} := \inf A \in A$, which implies that $\|\mathbf{X}\|_B \leq \|\mathbf{X}\|_{L^\infty}$ almost surely. Indeed, consider the sequence $\|\mathbf{X}\|_{L^\infty} + 1/j$, $j \in \mathbb{N}$. By the approximation property of the infimum, there exists a decreasing sequence of numbers $\{\rho_j\}_{j\in\mathbb{N}} \subset A$ in A satisfying $\|\mathbf{X}\|_{L^\infty} \leq \rho_j < \|\mathbf{X}\|_{L^\infty} + 1/j$ for all $j \in \mathbb{N}$. Define $F := \{\omega \in \Omega \mid \|\mathbf{X}(\omega)\|_B > \|\mathbf{X}\|_{L^\infty}\}$ and $F_j := \{\omega \in \Omega \mid \|\mathbf{X}(\omega)\|_B > \rho_j\}$. It is easy to see that $F_j \subset F_{j+1}$, $j \in \mathbb{N}$ and that $\lim_{j\to\infty} F_j = F$ and, consequently, (see Lemma 5, page 7 in Grimmett and Stirzaker (2001)) $\lim_{j\to\infty} \mathbb{P}(F_j) = \mathbb{P}(F)$. Since by construction $\mathbb{P}(F_j) = 0$ for all $j \in \mathbb{N}$ then $\mathbb{P}(F) = 0$ necessarily, which shows that $\|\mathbf{X}\|_{L^\infty} \in A$, as required.

(ii) If $\|\mathbf{X}\|_{L^\infty} \leq C$ then by part **(i)**, $\|\mathbf{X}\|_B \leq \|\mathbf{X}\|_{L^\infty} \leq C$ almost surely. Conversely, if $\|\mathbf{X}\|_B \leq C$ almost surely, then $C \in A = \left\{\rho \in \overline{\mathbb{R}^+} \mid \|\mathbf{X}\|_B \leq \rho \quad \text{almost surely}\right\}$. Consequently, $\|\mathbf{X}\|_{L^\infty} = \inf A \leq C \in A$, as required.

(iii) Suppose first that $\|\mathbf{X}\|_B \leq C$ almost surely and define $F := \{\omega \in \Omega \mid \|\mathbf{X}(\omega)\|_B > C\}$. By hypothesis, we have that $\mathbb{P}(F) = 0$ and $\mathbb{P}(\Omega \setminus F) = 1$. Then,

$$\begin{aligned}
\mathrm{E}\left[\|\mathbf{X}\|_B^k\right] &= \int_\Omega \|\mathbf{X}\|_B^k \, d\mathbb{P} = \int_{\Omega\setminus F} \|\mathbf{X}\|_B^k \, d\mathbb{P} + \int_F \|\mathbf{X}\|_B^k \, d\mathbb{P} \\
&= \int_{\Omega\setminus F} \|\mathbf{X}\|_B^k \, d\mathbb{P} \leq \int_{\Omega\setminus F} C^k d\mathbb{P} = C^k \mathbb{P}(\Omega \setminus F) = C^k,
\end{aligned}$$

as required. Conversely, assume that $\mathrm{E}\left[\|\mathbf{X}\|_B^k\right] \leq C^k$, for any $k \in \mathbb{N}$, and define

$$F_n := \left\{\omega \in \Omega \mid \|\mathbf{X}(\omega)\|_B > C + \frac{1}{n}\right\},$$

for all $n \geq 1$. It is easy to see that $F_n \subset F_{n+1}$ and that $\lim_{n\to\infty} F_n = F$ and, consequently, (see Lemma 5, page 7 in Grimmett and Stirzaker (2001)) $\lim_{n\to\infty} \mathbb{P}(F_n) = \mathbb{P}(F)$. Now,

$$\begin{aligned}
C^k &\geq \mathrm{E}\left[\|\mathbf{X}\|_B^k\right] = \int_\Omega \|\mathbf{X}\|_B^k \, d\mathbb{P} = \int_{\Omega\setminus F_n} \|\mathbf{X}\|_B^k \, d\mathbb{P} + \int_{F_n} \|\mathbf{X}\|_B^k \, d\mathbb{P} \\
&\geq \int_{F_n} \|\mathbf{X}\|_B^k \, d\mathbb{P} \geq \int_{F_n} \left(C + \frac{1}{n}\right)^k d\mathbb{P} = \left(C + \frac{1}{n}\right)^k \mathbb{P}(F_n),
\end{aligned}$$

which implies that $\mathbb{P}(F_n) \leq C^k / \left(C + \frac{1}{n}\right)^k$ for any $k \in \mathbb{N}$ and hence, by taking the limit $k \to \infty$, we can conclude that $\mathbb{P}(F_n) = 0$. Consequently, $\mathbb{P}(F) = \lim_{n\to\infty} \mathbb{P}(F_n) = 0$, which shows that $\|\mathbf{X}\|_B \leq C$ almost surely.

(iv) Let $||\cdot||$ denote the Euclidean norm on $\mathbb{R}^n$. Since $|X_i| \leq \|\mathbf{X}\|$ always and by part **(i)** $\|\mathbf{X}\| \leq \|\mathbf{X}\|_{L^\infty}$ almost surely, we can conclude that $|X_i| \leq \|\mathbf{X}\|_{L^\infty}$ almost surely. This implies that $X_i \in L^\infty(\Omega, \mathbb{R})$ and hence the statement follows from part **(iii)**. ■

6.12 Proof of Lemma 25

We start by proving by contradiction that

$$\operatorname*{ess\,sup}_{\omega\in\Omega} \left\{ \sup_{t\in\mathbb{Z}} \{\|\mathbf{z}_t(\omega)\|\} \right\} \geq \sup_{t\in\mathbb{Z}} \left\{ \operatorname*{ess\,sup}_{\omega\in\Omega} \{\|\mathbf{z}_t(\omega)\|\} \right\}. \tag{6.20}$$

Indeed, suppose that

$$\operatorname*{ess\,sup}_{\omega\in\Omega} \left\{ \sup_{t\in\mathbb{Z}} \{\|\mathbf{z}_t(\omega)\|\} \right\} < \sup_{t\in\mathbb{Z}} \left\{ \operatorname*{ess\,sup}_{\omega\in\Omega} \{\|\mathbf{z}_t(\omega)\|\} \right\}. \tag{6.21}$$

By the approximation property of the supremum (see Theorem 1.14 in Apostol (1974)), there exists $t_0 \in \mathbb{Z}$ such that

$$\operatorname*{ess\,sup}_{\omega\in\Omega}\left\{\sup_{t\in\mathbb{Z}}\left\{\|\mathbf{z}_t(\omega)\|\right\}\right\} < \operatorname*{ess\,sup}_{\omega\in\Omega}\left\{\|\mathbf{z}_{t_0}(\omega)\|\right\} \leq \sup_{t\in\mathbb{Z}}\left\{\operatorname*{ess\,sup}_{\omega\in\Omega}\left\{\|\mathbf{z}_t(\omega)\|\right\}\right\}. \tag{6.22}$$

However, $\|\mathbf{z}_{t_0}(\omega)\| \leq \sup_{t\in\mathbb{Z}}\{\|\mathbf{z}_t(\omega)\|\}$ for all $\omega\in\Omega$ and hence by part **(i)** in Lemma 24

$$\|\mathbf{z}_{t_0}(\omega)\| \leq \sup_{t\in\mathbb{Z}}\{\|\mathbf{z}_t(\omega)\|\} \leq \operatorname*{ess\,sup}_{\omega\in\Omega}\left\{\sup_{t\in\mathbb{Z}}\left\{\|\mathbf{z}_t(\omega)\|\right\}\right\}, \quad \text{almost surely.}$$

Now, by part **(ii)** in Lemma 24, this implies that

$$\operatorname*{ess\,sup}_{\omega\in\Omega}\left\{\|\mathbf{z}_{t_0}(\omega)\|\right\} \leq \operatorname*{ess\,sup}_{\omega\in\Omega}\left\{\sup_{t\in\mathbb{Z}}\left\{\|\mathbf{z}_t(\omega)\|\right\}\right\}.$$

However, this expression is in contradiction with the first inequality in (6.22) and hence the assumption (6.21) cannot be correct. This argument implies that the inequality (6.20) holds.

We now prove the reverse inequality, that is,

$$\operatorname*{ess\,sup}_{\omega\in\Omega}\left\{\sup_{t\in\mathbb{Z}}\left\{\|\mathbf{z}_t(\omega)\|\right\}\right\} \leq \sup_{t\in\mathbb{Z}}\left\{\operatorname*{ess\,sup}_{\omega\in\Omega}\left\{\|\mathbf{z}_t(\omega)\|\right\}\right\}. \tag{6.23}$$

By part **(ii)** of Lemma 24, this inequality holds if and only if

$$\sup_{t\in\mathbb{Z}}\left\{\|\mathbf{z}_t(\omega)\|\right\} \leq \sup_{t\in\mathbb{Z}}\left\{\operatorname*{ess\,sup}_{\omega\in\Omega}\left\{\|\mathbf{z}_t(\omega)\|\right\}\right\}, \quad \text{almost surely.} \tag{6.24}$$

Now, by part **(i)** in Lemma 24, we have that $\|\mathbf{z}_t(\omega)\| \leq \operatorname{ess\,sup}_{\omega\in\Omega}\left\{\|\mathbf{z}_t(\omega)\|\right\}$, almost surely and for each fixed $t\in\mathbb{Z}$. Let $A_t \subset \Omega$ be the zero-measure set such that $\|\mathbf{z}_t(\omega)\| > \operatorname{ess\,sup}_{\omega\in\Omega}\left\{\|\mathbf{z}_t(\omega)\|\right\}$ for all $\omega \in A_t$. Let $A := \bigcup_{t\in\mathbb{Z}} A_t$. Notice that $\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{t\in\mathbb{Z}} A_t\right) \leq \sum_{t\in\mathbb{Z}} \mathbb{P}(A_t) = 0$ and hence $B := A^{\mathsf{c}}$ has measure one and

$$\|\mathbf{z}_t(\omega)\| \leq \operatorname*{ess\,sup}_{\omega\in\Omega}\left\{\|\mathbf{z}_t(\omega)\|\right\}, \quad \text{for all } \omega\in B \text{ and all } t\in\mathbb{Z}.$$

Since B has measure one, this inequality is equivalent to (6.24), which guarantees that (6.23) holds. The inequalities (6.20) and (6.23) that we just proved imply that the equality (4.7) holds true. ■

6.13 Proof of Lemma 26

It is obvious that $S_{\ell^\infty(\mathbb{R}^n)} \subset S_{(\mathbb{R}^n)^{\mathbb{Z}}}$ and hence the inclusion map

$$\iota : S_{\ell^\infty(\mathbb{R}^n)} \hookrightarrow S_{(\mathbb{R}^n)^{\mathbb{Z}}}, \tag{6.25}$$

is well-defined. The equivariance with respect to the equivalence relations $\sim_{\ell^\infty(\mathbb{R}^n)}$ and $\sim_{(\mathbb{R}^n)^{\mathbb{Z}}}$ follows trivially from noticing that if $\mathbf{z}_1, \mathbf{z}_2 \in S_{\ell^\infty(\mathbb{R}^n)}$ are such that $\mathbf{z}_1 \sim_{\ell^\infty(\mathbb{R}^n)} \mathbf{z}_2$ one obviously have that $\iota(\mathbf{z})_1 \sim_{(\mathbb{R}^n)^{\mathbb{Z}}} \iota(\mathbf{z}_2)$. This shows the existence of the projected map ϕ that makes the diagram

$$\begin{array}{ccc} S_{\ell^\infty(\mathbb{R}^n)} & \overset{\iota}{\hookrightarrow} & S_{(\mathbb{R}^n)^{\mathbb{Z}}} \\ {\scriptstyle \Pi_{\sim_{\ell^\infty(\mathbb{R}^n)}}}\downarrow & & \downarrow{\scriptstyle \Pi_{\sim_{(\mathbb{R}^n)^{\mathbb{Z}}}}} \\ L^\infty\left(\Omega, \ell^\infty(\mathbb{R}^n)\right) & \overset{\phi}{\longrightarrow} & L^\infty\left(\Omega, (\mathbb{R}^n)^{\mathbb{Z}}\right), \end{array}$$

commutative where $\Pi_{\sim_{\ell^\infty(\mathbb{R}^n)}}$ and $\Pi_{\sim_{(\mathbb{R}^n)^{\mathbb{Z}}}}$ map the elements in $S_{\ell^\infty(\mathbb{R}^n)}$ and $S_{(\mathbb{R}^n)^{\mathbb{Z}}}$ onto their corresponding equivalence classes with respect to the associated equivalence relations. One can easily prove that the norm preservation following the diagram. It is a straightforward exercise to verify that ϕ is injective and preserves the norm $\|\cdot\|_{L^\infty}$. In order to show that ϕ is surjective, let $\mathbf{z} \in L^\infty\left(\Omega, (\mathbb{R}^n)^{\mathbb{Z}}\right)$. Given that $\|\mathbf{z}\|_{L^\infty} < \infty$ or, equivalently, $\operatorname{ess\,sup}_{\omega\in\Omega}\left\{\sup_{t\in\mathbb{Z}}\left\{\|\mathbf{z}_t(\omega)\|\right\}\right\} < \infty$, by part **(i)** in Lemma 24, this implies that

$$\sup_{t\in\mathbb{Z}}\{\|\mathbf{z}_t(\omega)\|\} < \infty, \quad \text{almost surely.} \tag{6.26}$$

Since the elements in the spaces in $L^\infty\left(\Omega, \ell^\infty(\mathbb{R}^n)\right)$ and $L^\infty\left(\Omega, (\mathbb{R}^n)^{\mathbb{Z}}\right)$ are equivalence classes containing almost surely equal random variables, we can take another representative $\mathbf{z}^* : \Omega \longrightarrow (\mathbb{R}^n)^{\mathbb{Z}}$ for the class containing $\mathbf{z} \in L^\infty\left(\Omega, (\mathbb{R}^n)^{\mathbb{Z}}\right)$ defined as

$$\mathbf{z}^*(\omega) := \begin{cases} \mathbf{z}(\omega), & \text{when} \quad \sup_{t\in\mathbb{Z}}\{\|\mathbf{z}_t(\omega)\|\} < \infty, \\ 0, & \text{otherwise.} \end{cases}$$

Since the processes $\mathbf{z}$ and $\mathbf{z}^*$ differ by (6.26) only in a set of zero measure, they are equal in $L^\infty\left(\Omega, (\mathbb{R}^n)^{\mathbb{Z}}\right)$ but, this time, $\mathbf{z}^* \in L^\infty\left(\Omega, \ell^\infty(\mathbb{R}^n)\right)$ and $\phi(\mathbf{z}^*) = \mathbf{z}$, as required. ■

6.14 Proof of Theorem 27

Proof of part (i). All along this proof we will denote the elements in K_M with a lower bold case ($\mathbf{z} \in K_M$) and those in $K_M^{L^\infty}$ with an upper bold case ($\mathbf{Z} \in K_M^{L^\infty}$).

We first assume that the functional $H : (K_M, \|\cdot\|_w) \longrightarrow \mathbb{R}$ has the fading memory property. This means that H is a continuous map and since by Lemma 2 the space $(K_M, \|\cdot\|_w)$ is compact, then so is the image $H(K_M)$ as a subset of the real line. This implies that there exists a finite real number $L > 0$ such that $H(K_M) \subset [-L, L]$. Let now $\mathbf{Z} \in K_M^{L^\infty}$; the condition $\|\mathbf{Z}\|_{L^\infty} \leq M$ is equivalent to $\|\mathbf{Z}_t\| \leq M$, for all $t \in \mathbb{Z}_-$, almost surely, and hence implies that $H(\mathbf{Z}) \in [-L, L]$, almost surely or, equivalently, that $\|H(\mathbf{Z})\|_{L^\infty} \leq L$. This, in turn, implies that $H(\mathbf{Z}) \in L^\infty(\Omega, \mathbb{R})$ for any $\mathbf{Z} \in K_M^{L^\infty}$, as required.

We now show that $H : (K_M^{L^\infty}, \|\cdot\|_{L_w^\infty}) \longrightarrow L^\infty(\Omega, \mathbb{R})$ has the FMP. The FMP hypothesis on $H : (K_M, \|\cdot\|_w) \longrightarrow \mathbb{R}$ implies that for any $\mathbf{z} \in K_M$ and any $\epsilon > 0$ there exists a $\delta(\epsilon) > 0$ such that for any $\mathbf{s} \in K_M$ that satisfies that

$$\|\mathbf{z} - \mathbf{s}\|_w = \sup_{t\in\mathbb{Z}_-}\{\|(\mathbf{z}_t - \mathbf{s}_t)w_{-t}\|\} < \delta(\epsilon), \quad \text{then} \quad |H(\mathbf{z}) - H(\mathbf{s})| < \epsilon. \tag{6.27}$$

Moreover, since by Lemma 2 the space $(K_M, \|\cdot\|_w)$ is compact, the Uniform Continuity Theorem (Theorem 7.3 in Munkres (2014)) guarantees that the relation $\delta(\epsilon)$ does not depend on the point $\mathbf{z} \in K_M$.

We now prove the statement by showing that for any $\epsilon > 0$ and $\mathbf{Z} \in K_M^{L^\infty}$ then $\|H(\mathbf{Z}) - H(\mathbf{S})\|_{L^\infty} < \epsilon$, for all $\mathbf{S} \in K_M^{L^\infty}$ such that $\|\mathbf{Z} - \mathbf{S}\|_{L_w^\infty} < \delta(\epsilon)$. Indeed, the inequality $\|\mathbf{Z} - \mathbf{S}\|_{L_w^\infty} < \delta(\epsilon)$ holds if and only if $\sup_{t\in\mathbb{Z}_-}\{\|\mathbf{Z}_t - \mathbf{S}_t\|_{L^\infty} w_{-t}\} < \delta(\epsilon)$. Given that for any $l \in \mathbb{Z}_-$ we have that $\|\mathbf{Z}_l - \mathbf{S}_l\|_{L^\infty} w_{-l} \leq \sup_{t\in\mathbb{Z}_-}\{\|\mathbf{Z}_t - \mathbf{S}_t\|_{L^\infty} w_{-t}\} < \delta(\epsilon)$, part **(ii)** in Lemma 24 implies that $\|\mathbf{Z}_l - \mathbf{S}_l\| w_{-l} < \delta(\epsilon)$ almost surely for any $l \in \mathbb{Z}_-$ and hence $\sup_{t\in\mathbb{Z}_-}\{\|\mathbf{Z}_t - \mathbf{S}_t\| w_{-t}\} = \|\mathbf{Z} - \mathbf{S}\|_w < \delta(\epsilon)$, almost surely. This implies, using (6.27), that $|H(\mathbf{Z}) - H(\mathbf{S})| < \epsilon$, almost surely, which by part **(ii)** in Lemma 24 implies that $\|H(\mathbf{Z}) - H(\mathbf{S})\|_{L^\infty} < \epsilon$, as required.

Conversely, if $H : (K_M^{L^\infty}, \|\cdot\|_{L_w^\infty}) \longrightarrow L^\infty(\Omega, \mathbb{R})$ has the fading memory property then so does $H : (K_M, \|\cdot\|_w) \longrightarrow \mathbb{R}$ because $K_M \subset K_M^{L^\infty}$ and $\|\mathbf{z}\| = \|\mathbf{z}\|_{L^\infty}$ for the elements $\mathbf{z} \in K_M$.

Proof of part (ii). We suppose first that $\mathcal{T}$ is dense in the set $(C^0(K_M), \|\cdot\|_w)$ and show that the corresponding family with intputs in $K_M^{L^\infty}$ is universal. Let $H : (K_M^{L^\infty}, \|\cdot\|_{L_w^\infty}) \longrightarrow L^\infty(\Omega, \mathbb{R})$ be an arbitrary causal and time-invariant FMP filter and let $H_S \in \mathcal{T}$ be such that $\sup_{\mathbf{z}\in K_M}\{\|H(\mathbf{z}) - H_S(\mathbf{z})\|_{L^\infty}\} < \epsilon$. The existence of H_S is ensured by the density hypothesis on $\mathcal{T}$. We show that this ensures that $\sup_{\mathbf{Z}\in K_M^{L^\infty}}\{\|H(\mathbf{Z})-H_S(\mathbf{Z})\|_{L^\infty}\}<\epsilon$. Indeed, this conclusion is true if $\|H(\mathbf{Z})-H_S(\mathbf{Z})\|_{L^\infty}<\epsilon$ for any $\mathbf{Z}\in K_M^{L^\infty}$ which, by part **(ii)** in Lemma 24 is equivalent to $|H(\mathbf{Z})-H_S(\mathbf{Z})|<\epsilon$ almost surely, for any $\mathbf{Z}\in K_M^{L^\infty}$. This condition is in turn true because as $\mathbf{Z}\in K_M^{L^\infty}$, then $\|\mathbf{Z}_t\|\leq M$ almost surely for all $t\in\mathbb{Z}_-$ and hence $\mathbf{Z}\in K_M$ almost surely. Since H_S approximates H for deterministic inputs, we have that $|H(\mathbf{Z})-H_S(\mathbf{Z})|<\epsilon$ almost surely, as required.

Conversely, if the family $\mathcal{T}$ with intputs in $K_M^{L^\infty}$ is universal in the set of continuous maps of the type $H:(K_M^{L^\infty},\|\cdot\|_{L_w^\infty})\longrightarrow L^\infty(\Omega,\mathbb{R})$ we can easily show that $\mathcal{T}$ is dense in $(C^0(K_M),\|\cdot\|_w)$. Let $H\in(C^0(K_M),\|\cdot\|_w)$ and let $H_S:(K_M^{L^\infty},\|\cdot\|_{L_w^\infty})\longrightarrow L^\infty(\Omega,\mathbb{R})$ be the element that, for a given $\epsilon>0$, satisfies $\|H-H_S\|_{L^\infty}=\sup_{\mathbf{Z}\in K_M^{L^\infty}}\{\|H(\mathbf{Z})-H_S(\mathbf{Z})\|_{L^\infty}\}<\epsilon$. Given that, as we pointed out, $K_M\subset K_M^{L^\infty}$ and $\|\mathbf{z}\|=\|\mathbf{z}\|_{L^\infty}$, for the elements $\mathbf{z}\in K_M$, we have

$$\|H-H_S\|=\sup_{\mathbf{z}\in K_M}\{\|H(\mathbf{z})-H_S(\mathbf{z})\|\}=\sup_{\mathbf{z}\in K_M}\{\|H(\mathbf{z})-H_S(\mathbf{z})\|_{L^\infty}\}\leq\sup_{\mathbf{Z}\in K_M^{L^\infty}}\{\|H(\mathbf{Z})-H_S(\mathbf{Z})\|_{L^\infty}\}<\epsilon. \quad \blacksquare$$

6.15 Proof of Lemma 28

As we pointed out in Section 2, if the reservoir system determined by $F:D_N\times\overline{B_n(\mathbf{0},M)}\longrightarrow D_N$ and $h:D_N\to\mathbb{R}$ has the echo state property, a result in Grigoryeva and Ortega (2018) guarantees that the associated filter is automatically causal and time-invariant. This implies the existence of a functional $H_h^F:(\mathbb{R}^n)^{\mathbb{Z}_-}\longrightarrow\mathbb{R}$ that, by hypothesis, has the fading memory property. The rest of the statement is a consequence of part **(i)** in Theorem 27. ■

6.16 Proof of Theorem 29

We first notice that the polynomial algebra $\mathcal{A}(\mathcal{R})$ is, by Theorem 8 and the first part of Theorem 27, made of fading memory reservoir filters that map into $L^\infty(\Omega,\mathbb{R})$. Using the other hypotheses in the statement we can easily conclude that the family $\mathcal{A}(\mathcal{R})$ satisfies the thesis of Theorem 8 and it is hence universal in the deterministic setup. The result follows from the second part of Theorem 27. ■

Acknowledgments: We thank Philipp Harms and Herbert Jaeger for carefully looking at early versions of this work and for making suggestions that have significantly improved some of our results. We thank Josef Teichmann for fruitful discussions. We also thank the editor and two remarkable anonymous referees whose input has significantly improved the presentation and the contents of the paper. The authors acknowledge partial financial support of the French ANR "BIPHOPROC" project (ANR-14-OHRI-0002-02) as well as the hospitality of the Centre Interfacultaire Bernoulli of the Ecole Polytechnique Fédérale de Lausanne during the program "Stochastic Dynamical Models in Mathematical Finance, Econometrics, and Actuarial Sciences" that made possible the collaboration that led to some of the results included in this paper. LG acknowledges partial financial support of the Graduate School of Decision Sciences and the Young Scholar Fund AFF of the Universität Konstanz. JPO acknowledges partial financial support coming from the Research Commission of the Universität Sankt Gallen and the Swiss National Science Foundation (grant number 200021_175801/1).

Glossary of Symbols

$\ell_w^\infty(\mathbb{R}^n)$	Banach space of semi-infinite sequences with finite weighted norm
$\mathbb{D}$	Space of diagonal matrices of any order
$\mathbb{D}_n$	Space of diagonal matrices of order $n\in\mathbb{N}$
$\mathbb{M}_n$	Space of square matrices of order $n\in\mathbb{N}$
$\mathbb{M}_{m,n}[z]$	$\mathbb{M}_{m,n}$-valued polynomials on z with coefficients in $\mathbb{M}_{m,n}$

$\mathbb{M}_{n,m}$	Space of real $n \times m$ matrices with $m, n \in \mathbb{N}$
$\mathbb{N}\mathrm{il}$	Space of nilpotent matrices of any order and any index
$\mathbb{N}\mathrm{il}[z]$	Space of matrix-valued nilpotent polynomials on z of any order and any index
$\mathbb{N}\mathrm{il}_n^k$	Space of nilpotent matrices of index $k \in \mathbb{N}$ in $\mathbb{M}_n$
$\mathbb{N}\mathrm{il}_n^k[z]$	Space of nilpotent $\mathbb{M}_n$-valued polynomials on z with coefficients in $\mathbb{M}_n$ of index k
$\mathcal{A}(\mathcal{R})$	Polynomial algebra generated by the set $\mathcal{R}$ of reservoir filters defined on K_M
$\mathcal{DL}_\epsilon$	Set of linear reservoir systems determined by diagonal matrices $A \in \mathbb{D}$ such that $\sigma_{\max}(A) < 1 - \epsilon$
$\mathcal{L}_\epsilon$	Set of linear reservoir systems determined by matrices $A \in \mathbb{M}_N$ such that $\sigma_{\max}(A) < 1 - \epsilon$
$\mathcal{NL}$	Set of linear reservoir systems determined by nilpotent matrices $A \in \mathbb{N}\mathrm{il}$
$\mathcal{NS}_\epsilon$	Subfamily of $\mathcal{S}_\epsilon$ formed by SAS reservoir systems determined by nilpotent polynomials p
$\mathcal{R}$	Set of reservoir filters defined on K_M
$\mathcal{S}_\epsilon$	State affine reservoir systems (SAS) $H^{p,q}_{\mathbf{W}} : I^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ with $M_p < 1-\epsilon$ and $M_q < 1-\epsilon$
$B_n(\mathbf{0}, M)$	Ball of radius M and center $\mathbf{0}$ in $\mathbb{R}^n$ with respect to the Euclidean norm
$F : \mathbb{R}^N \times \mathbb{R}^n \longrightarrow \mathbb{R}^N$	Reservoir map
$H_U : (\mathbb{R}^n)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$	Functional associated to the causal and time-invariant filter $U : (\mathbb{R}^n)^{\mathbb{Z}} \longrightarrow \mathbb{R}^{\mathbb{Z}}$
$h : \mathbb{R}^N \to \mathbb{R}$	Generic readout map
$H_h^{A,\mathbf{c}} : K_M \longrightarrow \mathbb{R}$	Linear reservoir functional determined by $A, \mathbf{c}$, and the polynomial h
$H^{p,q}_{\mathbf{W}} : I^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$	SAS reservoir functional
K_M	Space of semi-infinite sequences that are uniformly bounded by M
$K_M^{L^\infty}$	Space of semi-infinite processes that are almost surely uniformly bounded by M
$L^\infty\left(\Omega, (\mathbb{R}^n)^{\mathbb{Z}}\right)$	Space of almost surely bounded time series or discrete-time stochastic processes with values in $\mathbb{R}^n$
$L^\infty_w\left(\Omega, (\mathbb{R}^n)^{\mathbb{Z}_-}\right)$	Space of time series or discrete-time stochastic processes with values in $\mathbb{R}^n$ with finite L^∞_w-norm
N	Number of virtual neurons. Dimension of the reservoir state vectors
n	Dimension of the elements of the input signal
$U_h^F : (\mathbb{R}^n)^{\mathbb{Z}} \longrightarrow \mathbb{R}^{\mathbb{Z}}$	Reservoir filter
$U^F : (\mathbb{R}^n)^{\mathbb{Z}} \longrightarrow (\mathbb{R}^N)^{\mathbb{Z}}$	Filter determined by the reservoir map F
$U : (\mathbb{R}^n)^{\mathbb{Z}} \longrightarrow \mathbb{R}^{\mathbb{Z}}$	Filter with inputs in $\mathbb{R}^n$ and outputs in $\mathbb{R}$
$U_h^{A,\mathbf{c}} : K_M \longrightarrow \mathbb{R}^{\mathbb{Z}}$	Linear reservoir filter determined by $A, \mathbf{c}$, and the polynomial h
$U_H : (\mathbb{R}^n)^{\mathbb{Z}} \longrightarrow \mathbb{R}^{\mathbb{Z}}$	Causal and time-invariant filter associated to the functional $H : (\mathbb{R}^n)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$
$U^{p,q}_{\mathbf{W}} : I^{\mathbb{Z}} \longrightarrow \mathbb{R}^{\mathbb{Z}}$	SAS reservoir filter
$w : \mathbb{N} \longrightarrow (0, 1]$	Weighting sequence
$\mathbf{x}$	(Semi)-infinite sequence containing the reservoir states. The elements of this sequence are denoted by $\mathbf{x}_t \in \mathbb{R}^N$
$\mathbf{y}$	(Semi)-infinite output signal. The elements of this sequence are denoted by $y_t \in \mathbb{R}$
$\mathbf{z}$	(Semi)-infinite input signal. The elements of this sequence are denoted by $\mathbf{z}_t \in \mathbb{R}^n$

References

R. Abraham, J. E. Marsden, and T. S. Ratiu. *Manifolds, Tensor Analysis, and Applications*, volume 75. Applied Mathematical Sciences. Springer-Verlag, 1988.

T. Apostol. *Mathematical Analysis*. Addison Wesley, second edition, 1974.

L. Appeltant, M. C. Soriano, G. Van der Sande, J. Danckaert, S. Massar, J. Dambre, B. Schrauwen, C. R. Mirasso, and I. Fischer. Information processing using a single dynamical node as complex system. *Nature Communications*, 2:468, jan 2011.

V. I. Arnold. On functions of three variables. *Proceedings of the USSR Academy of Sciences*, 114: 679–681, 1957.

A. F. Atiya and A. G. Parlos. New results on recurrent network training: unifying the algorithms and accelerating convergence. *IEEE Transactions on Neural Networks*, 11(3):697–709, jan 2000.

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*, 23(1):175–182, jan 2012.

A. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, may 1993.

T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31 (3):307–327, 1986.

G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1976.

S. Boyd and L. Chua. Fading memory and the problem of approximating nonlinear operators with Volterra series. *IEEE Transactions on Circuits and Systems*, 32(11):1150–1161, nov 1985.

M. B. Brilliant. Theory of the analysis of nonlinear systems. Technical report, Massachusetts Institute of Technology, Research Laboratory of Electronics, 1958.

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

J. W. Brown and R. V. Churchill. *Complex Variables and Applications Eighth Edition*. McGraw-Hill, eighth edition, 2009.

D. Brunner, M. C. Soriano, C. R. Mirasso, and I. Fischer. Parallel photonic information processing at gigabyte per second data rates using transient states. *Nature Communications*, 4(1364), 2013.

M. Buehner and P. Young. A tighter bound for the echo state property. *IEEE Transactions on Neural Networks*, 17(3):820–824, 2006.

F. Comets and T. Meyre. *Calcul Stochastique et Modèles de Diffusions*. Dunod, Paris, 2006.

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

N. Crook. Nonlinear transient computation. *Neurocomputing*, 70:1167–1176, 2007.

J. P. Crutchfield, W. L. Ditto, and S. Sinha. Introduction to focus issue: intrinsic and designed computation: information processing in dynamical systems-beyond the digital hegemony. *Chaos*, 20(3): 037101, sep 2010.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, dec 1989.

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

J. Dieudonné. *Foundations of Modern Analysis.* Academic Press, 1969.

K. Doya. Bifurcations in the learning of recurrent neural networks. In *Proceedings of IEEE International Symposium on Circuits and Systems*, volume 6, pages 2777–2780. IEEE, 1992.

R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4):987–1007, 1982.

M. Fliess. Un outil algebrique : les series formelles non commutatives. In G. Marchesini and S. K. Mitter, editors, *Mathematical Systems Theory*, pages 122–148. Springer Verlag, 1976.

M. Fliess and D. Normand-Cyrot. Vers une approche algébrique des systèmes non linéaires en temps discret. In A. Bensoussan and J. Lions, editors, *Analysis and Optimization of Systems. Lecture Notes in Control and Information Sciences, vol. 28.* Springer Berlin Heidelberg, 1980.

C. Francq and J.-M. Zakoian. *GARCH Models: Structure, Statistical Inference and Financial Applications.* Wiley, 2010.

M. Fréchet. Sur les fonctionnelles continues. *Annales scientifiques de l'Ecole Normale Supérieure. 3ème série.*, 27:193–216, 1910.

M. N. Galtier, C. Marini, G. Wainrib, and H. Jaeger. Relative entropy minimizing noisy non-linear neural network to approximate stochastic processes. *Neural Networks*, 56:10–21, 2014.

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

D. A. George. Continuous nonlinear systems. Technical report, Massachusetts Institute of Technology, Research Laboratory of Electronics, 1959.

L. Gonon and J.-P. Ortega. Reservoir computing universality with stochastic inputs. *Preprint*, 2018.

A. Graves, A.-R. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649. IEEE, may 2013.

L. Grigoryeva and J.-P. Ortega. Echo state networks are universal. *Preprint*, 2018.

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

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

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

G. Grimmett and D. Stirzaker. *Probability and Random Processes.* Oxford University Press, 2001.

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*, 23(3):341–55, apr 2010.

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

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

T. W. Hungerford. *Algebra*. Springer New York, 1974.

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

H. Jaeger. The 'echo state' approach to analysing and training recurrent neural networks with an erratum note. Technical report, German National Research Center for Information Technology, 2010.

H. Jaeger and H. Haas. Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication. *Science*, 304(5667):78–80, 2004.

H. Jaeger, M. Lukoševičius, D. Popovici, and U. Siewert. Optimization and applications of echo state networks with leaky-integrator neurons. *Neural Networks*, 20(3):335–352, 2007.

L. K. Jones. A simple lemma on greedy approximation in hilbert space and convergence rates for projection pursuit regression and neural network training. *The Annals of Statistics*, 20(1):608–613, 1992.

A. N. Kolmogorov. On the representation of continuous functions of several variables as superpositions of functions of smaller number of variables. *Soviet Math. Dokl*, 108:179–182, 1956.

V. Kurkova and M. Sanguineti. Learning with generalization capability by kernel methods of bounded complexity. *Journal of Complexity*, 21(3):350–367, 2005.

L. Larger, M. C. Soriano, D. Brunner, L. Appeltant, J. M. Gutierrez, L. Pesquera, C. R. Mirasso, and I. Fischer. Photonic information processing beyond Turing: an optoelectronic implementation of reservoir computing. *Optics Express*, 20(3):3241, jan 2012.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer-Verlag, 1991.

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

W. Maass. Liquid state machines: motivation, theory, and applications. In S. S. Barry Cooper and A. Sorbi, editors, *Computability In Context: Computation and Logic in the Real World*, chapter 8, pages 275–296. 2011.

W. Maass and E. D. Sontag. Neural Systems as Nonlinear Filters. *Neural Computation*, 12(8):1743–1772, aug 2000.

W. Maass, T. Natschläger, and H. Markram. Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Computation*, 14:2531–2560, 2002.

W. Maass, T. Natschläger, and H. Markram. Fading memory and kernel properties of generic cortical microcircuit models. *Journal of Physiology Paris*, 98(4-6 SPEC. ISS.):315–330, 2004.

W. Maass, P. Joshi, and E. D. Sontag. Computational aspects of feedback in neural circuits. *PLoS Computational Biology*, 3(1):e165, 2007.

G. Manjunath and H. Jaeger. Echo state property linked to an input: exploring a fundamental characteristic of recurrent neural networks. *Neural Computation*, 25(3):671–696, 2013.

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

M. B. Matthews. Approximating nonlinear fading-memory operators using neural network models. *Circuits, Systems, and Signal Processing*, 12(2):279–307, jun 1993.

J. Munkres. *Topology.* Pearson, second edition, 2014.

Y. Paquot, F. Duport, A. Smerieri, J. Dambre, B. Schrauwen, M. Haelterman, and S. Massar. Optoelectronic reservoir computing. *Scientific reports*, 2:287, jan 2012.

R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio. How to construct deep recurrent neural networks. *arXiv*, dec 2013.

J. Pathak, Z. Lu, B. R. Hunt, M. Girvan, and E. Ott. Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data. *Chaos*, 27(12), 2017.

J. Pathak, B. Hunt, M. Girvan, Z. Lu, and E. Ott. Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data: A Reservoir Computing Approach. *Physical Review Letters*, 120(2): 24102, 2018.

P. Perryman and A. Stubberud. Uniform, in-probability approximation of stochastic systems. In *Conference Record of The Thirtieth Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 146–150. IEEE Comput. Soc. Press, 1997.

P. C. Perryman. *Approximation Theory for Deterministic and Stochastic Nonlinear Systems.* PhD thesis, University of California, Irvine, 1996.

G. Pisier. Remarques sur un résultat non publié de B. Maurey. *Séminaire d'analyse fonctionnelle École Polytechnique*, pages 1–12, 1981.

G. Pisier. *Martingales in Banach Spaces.* Cambridge University Press, 2016.

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

L. Rüschendorf and W. Thomsen. Closedness of sum spaces and the generalized schrödinger Problem. *Theory of Probability & Its Applications*, 42(3):483–494, jan 1998.

E. Sontag. Realization theory of discrete-time nonlinear systems: Part I-The bounded case. *IEEE Transactions on Circuits and Systems*, 26(5):342–356, may 1979a.

E. D. Sontag. Polynomial Response Maps. In *Lecture Notes Control in Control and Information Sciences. Vol. 13.* Springer Verlag, 1979b.

D. A. Sprecher. A representation theorem for continuous functions of several variables. *Proceedings of the American Mathematical Society*, 16(2):200, apr 1965.

D. A. Sprecher. A numerical implementation of Kolmogorov's superpositions. *Neural Networks*, 9(5): 765–772, 1996.

D. A. Sprecher. A numerical implementation of Kolmogorov's superpositions II. *Neural Networks*, 10 (3):447–457, 1997.

A. Stubberud and P. Perryman. State of system approximation for stochastic systems. In *Proceedings of 13th International Conference on Digital Signal Processing*, volume 2, pages 711–714. IEEE, 1997a.

A. Stubberud and P. Perryman. Current state of system approximation for deterministic and stochastic systems. In *Conference Record of The Thirtieth Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 141–145. IEEE Comput. Soc. Press, 1997b.

H. J. Sussmann. Semigroup representations, bilinear approximations of input-output maps, and generalized inputs. In G. Marchesini and S. K. MItter, editors, *Mathematical Systems Theory*, pages 172–191. Springer Verlag, 1976.

F. Takens. Detecting strange attractors in turbulence. pages 366–381. Springer Berlin Heidelberg, 1981.

K. Vandoorne, J. Dambre, D. Verstraeten, B. Schrauwen, and P. Bienstman. Parallel reservoir computing using optical amplifiers. *IEEE Transactions on Neural Networks*, 22(9):1469–1481, sep 2011.

K. Vandoorne, P. Mechet, T. Van Vaerenbergh, M. Fiers, G. Morthier, D. Verstraeten, B. Schrauwen, J. Dambre, and P. Bienstman. Experimental demonstration of reservoir computing on a silicon photonics chip. *Nature Communications*, 5:78–80, mar 2014.

D. Verstraeten, B. Schrauwen, M. D'Haene, and D. Stroobandt. An experimental unification of reservoir computing methods. *Neural Networks*, 20:391–403, 2007.

Q. Vinckier, F. Duport, A. Smerieri, K. Vandoorne, P. Bienstman, M. Haelterman, and S. Massar. High-performance photonic reservoir computer based on a coherently driven passive cavity. *Optica*, 2(5):438–446, 2015.

G. Wainrib and M. N. Galtier. A local echo state property through the largest Lyapunov exponent. *Neural Networks*, 76:39–45, apr 2016.

O. White, D. Lee, and H. Sompolinsky. Short-Term Memory in Orthogonal Neural Networks. *Physical Review Letters*, 92(14):148102, apr 2004.

N. Wiener. *Nonlinear Problems in Random Theory.* The Technology Press of MIT, 1958.

I. B. Yildiz, H. Jaeger, and S. J. Kiebel. Re-visiting the echo state property. *Neural Networks*, 35:1–9, nov 2012.

G. Zang and P. A. Iglesias. Fading memory and stability. *Journal of the Franklin Institute*, 340(6-7): 489–502, 2004.

W. Zaremba, I. Sutskever, and O. Vinyals. Recurrent neural network regularization. *arXiv*, sep 2014.