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

Lyudmila Grigoryeva[1] and Juan-Pablo Ortega[2,3]

**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.

**Key Words:** reservoir computing, state-affine systems, SAS, echo state networks, ESN, echo state affine systems, ESAS, machine learning, universality, fading memory property, linear training, stochastic signal treatment.

## 1 Introduction

A ***reservoir computer (RC)*** [Jaeg 10, Jaeg 04, Maas 02, Maas 11, Croo 07, Vers 07, Luko 09] 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.

In supervised machine learning applications, the reservoir map is very often randomly generated (see, for instance, the echo state networks in [Jaeg 10, Jaeg 04]) and the memoryless readout is trained

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

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

[3]Centre National de la Recherche Scientifique (CNRS). France.

so that the output matches a given ***teaching signal*** that we will 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 attained with readouts that have a relatively simple functional form. Indeed, in many occasions it suffices to use an affine map $h : \mathbb{R}^N \to \mathbb{R}$ that is 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 that render classical gradient descent methods non-convergent [Doya 92].

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 the estimation of the error that is going to be committed in the execution of a specific task. This estimation can have the form of generic bounds that incorporate various architecture parameters of the system like in [Pisi 81, Jone 92, Barr 93, Kurk 05]. In the specific context of reservoir computing, it has been the subject of much research [Jaeg 02, Whit 04, Gang 08, Herm 10, Damb 12, Grig 15, Coui 16, Grig 16a].

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 all measurable functionals. For classical machine learning paradigms like neural networks, this question has given rise to well-known results [Kolm 56, Arno 57, Spre 65, Spre 96, Spre 97, Cybe 89, Horn 89, Rusc 98] that show that they can be considered as universal approximators in a static and deterministic setup.

There is no general recipe that allows one to conclude the universality of a given supervised machine learning paradigm. The proof strategy depends much on the specific paradigm itself 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, that is, when dealing with real-valued curves or time series. 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 is obtained as a consequence of various versions of the Stone-Weierstrass, that is 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 is obtained by restricting the study to functionals that satisfy a condition introduced in [Boyd 85] and 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 obtained as a corollary of classical results in systems theory. Indeed, in the continuous time setup, it can be established [Flie 76, Suss 76] for linear systems using polynomial readouts or with bilinear systems using linear readouts. In the discrete-time setup, the situation is more convoluted when the readout is linear and required the introduction in [Flie 80] of the so-called (homogeneous) ***state-affine systems (SAS)*** (see also [Sont 79a, Sont 79b]). The extension of these results to continuous non-compact time intervals was carried out in [Boyd 85] for fading memory filters using exponentially stable linear RCs with polynomial readouts and their bilinear counterparts with linear readouts (see also [Maas 00, Maas 02, Maas 04, Maas 07]). An extension to the non-compact discrete-time setup based on the Stone-Weierstrass theorem is, to our knowledge, not available in the literature. 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 [Matt 92, Matt 93]

for absolutely summable systems. The proofs in those works were unfortunately based on an invalid compactness assumption. Even though corrections were proposed in [Perr 96, Stub 97a], 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.

Our paper contains the following four main contributions:

- A non-homogeneous variant of the state-affine systems in [Flie 80] is introduced and we identify sufficient conditions that guarantee that the associated reservoir computers with linear readouts have the echo state property, are causal, time-invariant, and satisfy the fading memory property.
- A subset of this class is characterized that is universal in the category of fading memory filters with uniformly bounded inputs.
- These results are extended to the stochastic setup by formulating a version of this universality result that is valid for almost surely uniformly bounded inputs. This result shows 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.
- The universal non-homogeneous state-affine family suggests a generalization of the echo state networks introduced in [Jaeg 04] that have been very successful in many information processing tasks. We call this generalization ***echo state affine systems (ESAS)*** and we empirically show that they outperform echo state networks in a standard benchmark forecasting task having to do with the chaotic time series generated by the Mackey-Glass system [Mack 77].

Despite some preexisting work on the uniform approximation in probability of stochastic systems with finite memory [Perr 96, Perr 97, Stub 97b], 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 [Galt 14]. Indeed, in the deterministic setup, RC has been very successful (see for instance [Jaeg 04]) at the time of learning the attractors of various dynamical systems which, in passing, is used for forecasting by continuing the paths of the system in question using its synthetically learnt proxy. This approach led to several orders of magnitude accuracy improvements with respect to most standard dynamical systems forecasting tecniques based on Takens' Theorem [Take 81] and we expect that should have analogous beneficial effects in the density forecasting of stochastic processes that satisfy the hypotheses of the results that are formulated later on in the paper.

## 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 $v_i$, with $i \in \{1, \ldots, n\}$; we also write $\mathbf{v} = (v_i)_{i\in\{1,\ldots,n\}}$. 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 $n$. Given a matrix $A \in \mathbb{M}_{n,m}$, we denote its components by $A_{ij}$ and we write $A = (A_{ij})$, with $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots m\}$. Given a vector $\mathbf{v} \in \mathbb{R}^n$, the symbol $\|\mathbf{v}\|$ stands for its Euclidean norm. For any $A \in \mathbb{M}_{n,m}$, $\|A\|_2$ denotes its matrix norm induced by the Euclidean norms in $\mathbb{R}^m$ and $\mathbb{R}^n$, and satisfies [Horn 13, Example 5.6.6] that $\|A\|_2 = \sigma_{\max}(A)$, with $\sigma_{\max}(A)$ the largest singular value of $A$. $\|A\|_2$ is sometimes referred to as the spectral norm of $A$ [Horn 13].

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 $z \in \mathbb{R}$

is a scalar 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

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

**Filters.** The symbol $(\mathbb{R}^n)^{\mathbb{Z}}$ denotes the set of infinite real sequences of the form $\mathbf{z} = (\ldots, \mathbf{z}_{-1}, \mathbf{z}_0, \mathbf{z}_1, \ldots)$, $\mathbf{z}_i \in \mathbb{R}^n$; $(\mathbb{R}^n)^{\mathbb{Z}_-}$ and $(\mathbb{R}^n)^{\mathbb{Z}_+}$ are the subspaces consisting of, respectively, left and right infinite sequences: $(\mathbb{R}^n)^{\mathbb{Z}_-} = \{\mathbf{z} = (\ldots \mathbf{z}_{-2}, \mathbf{z}_{-1}, \mathbf{z}_0) \mid \mathbf{z}_i \in \mathbb{R}^n\}$, $(\mathbb{R}^n)^{\mathbb{Z}_+} = \{\mathbf{z} = (\mathbf{z}_0, \mathbf{z}_1, \mathbf{z}_2, \ldots) \mid \mathbf{z}_i \in \mathbb{R}^n\}$. In most cases we shall use in these infinite product spaces either the product topology (see [Munk 14, Chapter 2]) or the topology induced by the supremum norm $\|\mathbf{z}\|_\infty := \sup_{n\in\mathbb{Z}}\{\ldots, \mathbf{z}_{-1}, \mathbf{z}_0, \mathbf{z}_1, \ldots\}$. 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 $\|\cdot\|_\infty$.

We will refer to the maps of the type $U : (\mathbb{R}^n)^{\mathbb{Z}} \longrightarrow \mathbb{R}^{\mathbb{Z}}$ as ***filters*** or ***operators*** and to those like $H : (\mathbb{R}^n)^{\mathbb{Z}} \longrightarrow \mathbb{R}$ (or $H : (\mathbb{R}^n)^{\mathbb{Z}_\pm} \longrightarrow \mathbb{R}$) as ***functionals***. A filter $U : (\mathbb{R}^n)^{\mathbb{Z}} \longrightarrow \mathbb{R}^{\mathbb{Z}}$ is called ***causal*** when for any two elements $\mathbf{z}, \mathbf{w} \in \mathbb{R}^{\mathbb{Z}}$ that satisfy that $\mathbf{z}_\tau = \mathbf{w}_\tau$ for any $\tau \leq t$, for a given $t \in \mathbb{Z}$, we have that $(U\mathbf{z})_t = (U\mathbf{w})_t$. The filter $U$ is called ***time-invariant*** when, for any $\tau \in \mathbb{Z}$, it commutes with the associated time delay operator $U_\tau : (\mathbb{R}^n)^{\mathbb{Z}} \longrightarrow (\mathbb{R}^n)^{\mathbb{Z}}$ defined by $(U_\tau \mathbf{z})_t := \mathbf{z}_{t-\tau}$, that is, $U_\tau \circ U = U \circ U_\tau$ (in this expression, the two time delay operators $U_\tau$ have to be understood as defined in the appropriate sequence spaces). We recall (see for instance [Boyd 85]) that there is a bijection between causal time-invariant filters and functionals on $(\mathbb{R}^n)^{\mathbb{Z}_-}$. Indeed, given a time-invariant filter $U$, we can associate to it a functional $H_U : (\mathbb{R}^n)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ via the assignment $H_U(\mathbf{z}) := U(\mathbf{z}^e)_0$, where $\mathbf{z}^e \in (\mathbb{R}^n)^{\mathbb{Z}}$ is an arbitrary extension of $\mathbf{z} \in (\mathbb{R}^n)^{\mathbb{Z}_-}$ to $(\mathbb{R}^n)^{\mathbb{Z}}$. Conversely, for any functional $H : (\mathbb{R}^n)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$, we can define a time-invariant causal filter $U_H : (\mathbb{R}^n)^{\mathbb{Z}} \longrightarrow (\mathbb{R}^n)^{\mathbb{Z}}$ 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 $\mathbb{P}_{\mathbb{Z}_-} : (\mathbb{R}^n)^{\mathbb{Z}} \longrightarrow (\mathbb{R}^n)^{\mathbb{Z}_-}$ is the natural projection. It is easy to verify that:

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

Additionally, let $H_1, H_2 : (\mathbb{R}^n)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ 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 $\mathbf{z} \in (\mathbb{R}^n)^{\mathbb{Z}}$.

**Reservoir filters.** Consider now the RC system determined by (1.1)–(1.2) and assume, additionally, that the following existence and uniqueness property holds: for each $\mathbf{z} \in (\mathbb{R}^n)^{\mathbb{Z}}$ there exists a unique $\mathbf{x} \in \left(\mathbb{R}^N\right)^{\mathbb{Z}}$ 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*** [Jaeg 10, Yild 12] and has deserved much attention in the context of echo state networks [Jaeg 04, Bueh 06, Bai 12, Wain 16, Manj 13]. We emphasize that the echo state property is a genuine condition that is not automatically satisfied by all RC systems.

RC systems that satisfy the echo state property have a naturally associated filter. We will denote by $U^F : (\mathbb{R}^n)^{\mathbb{Z}} \longrightarrow (\mathbb{R}^N)^{\mathbb{Z}}$ the filter determined by the reservoir map via (1.1), that is, $(U^F\mathbf{z})_t := \mathbf{x}_t \in \mathbb{R}^N$ and by $U^F_h : (\mathbb{R}^n)^{\mathbb{Z}} \longrightarrow \mathbb{R}^{\mathbb{Z}}$ the one determined by the entire reservoir system, that is, $(U^F_h\mathbf{z})_t := y_t$. $U^F_h$ 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 [Grig 18]. We can hence associate to $U^F_h$ a ***reservoir functional*** $H^F_h : (\mathbb{R}^n)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ 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 and $D \subset \mathbb{R}^n$. We define the associated ***weighted norm*** $\|\cdot\|_w$ on $D^{\mathbb{Z}_-} \subset (\mathbb{R}^n)^{\mathbb{Z}_-}$ associated to the ***weighting sequence*** $w$ as the map:

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

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

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

endowed with weighted norm $\|\cdot\|_w$ forms a Banach space [Grig 18].

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 2.1** *Let $M > 0$ and let $K_M$ be the set of uniformly bounded elements in $D^{\mathbb{Z}_-}$ by $M$, that is,*

$$K_M := \left\{ \mathbf{z} \in D^{\mathbb{Z}_-} \mid \|\mathbf{z}_t\| \leq M \quad \text{for all} \quad t \in \mathbb{Z}_- \right\}. \tag{2.3}$$

*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 following series are convergent and satisfy 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 85] 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 2.2** *Let $M > 0$ and let $K_M$ be as in (2.3). Let $w : \mathbb{N} \longrightarrow (0,1]$ be an arbitrary weighting sequence. Then $K_M$ is a compact topological space when endowed with the relative topology inherited from $(D^{\mathbb{Z}_-}, \|\cdot\|_w)$.*

**Definition 2.3** *Let $K \subset D^{\mathbb{Z}_-}$ and let $H_U : (\mathbb{R}^n)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ be the functional associated to the causal and time-invariant filter $U : (\mathbb{R}^n)^{\mathbb{Z}} \longrightarrow (\mathbb{R}^n)^{\mathbb{Z}}$. We say that $U$ has the **fading memory property (FMP)** on $K$ whenever there exists a weighting sequence $w : \mathbb{N} \longrightarrow (0,1]$ such that the map $H_U : K \subset (D^{\mathbb{Z}_-}, \|\cdot\|_w) \longrightarrow \mathbb{R}$ is continuous. This means that for any $\mathbf{z} \in K$ and any $\epsilon > 0$ there exists a $\delta(\epsilon) > 0$ such that for any $\mathbf{s} \in K$ that satisfies that*

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

*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 $\lambda$-**exponential fading memory property**.*

**Remark 2.4** The fading memory property is in some occasions also related to the ***Lyapunov stability*** of the autonomous system associated to the reservoir map. 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 04] or for time-delay reservoirs [Grig 16b].

**Remark 2.5** 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 2.3 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.

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

**Lemma 2.6** *Let $K \subset D^{\mathbb{Z}_-}$ and let $H_U : (\mathbb{R}^n)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ be the functional associated to the causal and time-invariant filter $U : (\mathbb{R}^n)^{\mathbb{Z}} \longrightarrow (\mathbb{R}^n)^{\mathbb{Z}}$. If $H_U : K \subset D^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ has the FMP with respect to a given weighting sequence $w$, then it also has it with respect to any other weighting sequence $w'$ that satisfies that*

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

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

It can be shown [Grig 18] that when in this lemma the set $K$ is made of uniformly bounded sequences, that is, $K = K_M$, with $K_M$ as in (2.1), 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

In this section we consider deterministic filters, in the sense that their inputs belong to a subset of $K_M \subset (\mathbb{R}^n)^{\mathbb{Z}}$ formed by uniformly bounded elements, as in the definition in (2.3).

We will formulate two different universality results. In the first one, we show that polynomial algebras of filters generated by reservoir computers with the fading memory property that separate points are able to approximate any fading memory filter. Such families of reservoir computers are said to be ***universal***. Two important consequences of this result is 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. 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 first result can be seen as a discrete-time translation of the one formulated in [Boyd 85] for continuous-time filters while the second one is an extension to infinite time intervals of the main approximation result in [Flie 80] 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 theorem for polynomial subalgebras of real-valued functions defined on compact metric spaces, as formulated in Theorem 7.3.1 in [Dieu 69]. See Appendix 6.4 for a detailed proof.

**Theorem 3.1** *Let $K_M \subset (\mathbb{R}^n)^{\mathbb{Z}_-}$ be a subset of the type defined in (2.3) and let*

$$\mathcal{R} := \{H_{h_i}^{F_i} : K_M \longrightarrow \mathbb{R} \mid h_i \in C^\infty(\mathbb{R}^{N_i}), F_i : \mathbb{R}^{N_i} \times \mathbb{R}^n \longrightarrow \mathbb{R}^{N_i}, i \in I, N_i \in \mathbb{N}\} \tag{3.1}$$

*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 $\mathcal{R}$. 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 indices $\{i_1, \ldots, i_r\} \subset I$ and a polynomial $p : \mathbb{R}^r \longrightarrow \mathbb{R}$ such that*

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

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 $h_i \in C^\infty(\mathbb{R}^{N_i})$, $F_i : \mathbb{R}^{N_i} \times \mathbb{R}^n \longrightarrow \mathbb{R}^{N_i}$, $i \in \{1,2\}$, 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 are such that

$$H^{F_1}_{h_1} \cdot H^{F_2}_{h_2} = H^F_h, \quad \text{with} \quad h := h_1 \cdot h_2 \in C^\infty(\mathbb{R}^{N_1+N_2}), \tag{3.2}$$

$$H^{F_1}_{h_1} + \lambda H^{F_2}_{h_2} = H^F_{h'}, \quad \text{with} \quad h' := h_1 + \lambda h_2 \in C^\infty(\mathbb{R}^{N_1+N_2}), \tag{3.3}$$

and where

$$F : \mathbb{R}^{N_1+N_2} \times \mathbb{R}^n \longrightarrow \mathbb{R}^{N_1+N_2} \quad \text{is the map given by} \quad F((\mathbf{x}_1,\mathbf{x}_2),\mathbf{z}) := (F_1(\mathbf{x}_1,\mathbf{z}), F_2(\mathbf{x}_2,\mathbf{z}))\,. \tag{3.4}$$

We emphasize that the functionals $H^F_h$ and $H^F_{h'}$ in (3.2) and (3.3) are well defined because if the reservoir maps $F_1$ and $F_2$ satisfy the echo state property then so does $F$. Indeed, if $\mathbf{x}_1 \in \left(\mathbb{R}^{N_1}\right)^{\mathbb{Z}}$ and $\mathbf{x}_2 \in \left(\mathbb{R}^{N_2}\right)^{\mathbb{Z}}$ are the solutions of the reservoir equation (1.1) for $F_1$ and $F_2$ associated to the input $\mathbf{z} \in (\mathbb{R}^n)^{\mathbb{Z}}$, then so is $\mathbf{x}_1 \oplus \mathbf{x}_2 \in \left(\mathbb{R}^{N_1+N_2}\right)^{\mathbb{Z}}$ for $F$ in (3.4).

This observation has as a consequence that the set of 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 $C^\infty(\mathbb{R}^{N_i})$) and separate points (take the trivial reservoir map $F(\mathbf{x},\mathbf{z}) = \mathbf{z}$ and use the separation property of $C^\infty(\mathbb{R}^n)$). This remark, put together with Theorem 3.1 yields the following corollary.

**Corollary 3.2** *Let $K_M \subset (\mathbb{R}^n)^{\mathbb{Z}_-}$ be a subset of the type defined in (2.3) and let*

$$\mathcal{R}_w := \{H^F_h : K_M \longrightarrow \mathbb{R} \mid h \in C^\infty(\mathbb{R}^N), F : \mathbb{R}^N \times \mathbb{R}^n \longrightarrow \mathbb{R}^N, N \in \mathbb{N}\} \tag{3.5}$$

*be the set of all reservoir filters defined on $K_M$ 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 $(K_M, \|\cdot\|_w)$.*

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

$$\begin{cases} \mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{c}\mathbf{z}_t, \quad A \in \mathbb{M}_N, \mathbf{c} \in \mathbb{M}_{N,n}, & (3.6) \\ y_t = h(\mathbf{x}_t), \quad h \in \mathbb{R}[\mathbf{x}]. & (3.7) \end{cases}$$

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

**Corollary 3.3** *Let $K_M \subset (\mathbb{R}^n)^{\mathbb{Z}_-}$ be a subset of the type defined in (2.3) and let $\epsilon \in \mathbb{R}$ be such that $1 > \epsilon > 0$. Consider the set $\mathcal{L}_\epsilon$ formed by all the linear reservoir systems as in (3.6)-(3.7) determined by polynomial readouts and by matrices $A \in \mathbb{M}_N$ such that $\sigma_{\max}(A) < 1 - \epsilon$. Then, the elements in $\mathcal{L}_\epsilon$ generate $\lambda_\rho$-exponential fading memory reservoir filters, with $\lambda_\rho := (1-\epsilon)^\rho$, for any $\rho \in (0,1)$, that can be explicitly written down using the expression:*

$$H^{A,\mathbf{c}}_h(\mathbf{z}) = h\left(\sum_{i=0}^{\infty} A^i \mathbf{c}\mathbf{z}_{-i}\right), \quad \textit{for any} \quad \mathbf{z} \in K_M. \tag{3.8}$$

*This family is dense in $(C^0(K_M), \|\cdot\|_{w^\rho})$ with $w^\rho_t := \lambda^t_\rho$, for any $\rho \in (0,1)$.*

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

As it was explained in the introduction, it is is particularly convenient to work with RCs that have a linear readout since in that case the training reduces to the solution of a regression problem. That is in most cases feasible when there is need, as it happens in many applications, for a high number of neurons. 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 3.12).

**Definition 3.4** *Let $N \in \mathbb{N}$, $\mathbf{W} \in \mathbb{R}^N$, and let $p(z) \in \mathbb{M}_{N,N}[z]$ 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:*

$$\begin{cases} \mathbf{x}_t = p(z_t)\mathbf{x}_{t-1} + q(z_t), & (3.9) \\ y_t = \mathbf{W}^\top \mathbf{x}_t. & (3.10) \end{cases}$$

Notice that the linear reservoir equation (3.6) is a particular case of the SAS state equation (3.9) that is obtained by taking for $p$ and $q$ polynomials of degree zero.

Our next result spells out a sufficient condition that guarantees that the SAS reservoir system (3.9)-(3.10) has the echo state property. Moreover, it provides an explicit expression for the unique causal and time-invariant solution associated to a given input. Recall that 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 that $\|A\|_2 = \sigma_{\max}(A)$.

**Proposition 3.5** *Consider a non-homogeneous state-affine system as in (3.9)-(3.10) determined by polynomials $p, q$, and a vector $\mathbf{W}$, with inputs defined on $I^{\mathbb{Z}}$, $I := [-1, 1]$. Assume that*

$$\max_{z \in I} \|p(z)\|_2 = \max_{z \in I} \sigma_{\max}(p(z)) < 1. \tag{3.11}$$

*Then, the reservoir system (3.9)-(3.10) has the echo state property and for each input $\mathbf{z} \in \mathbb{R}^{\mathbb{Z}}$ it has a unique causal and time-invariant solution given by*

$$\begin{cases} \mathbf{x}_t = \displaystyle\sum_{j=0}^{\infty} \left( \prod_{k=0}^{j-1} p(z_{t-k}) \right) q(z_{t-j}), & (3.12) \\ y_t = \mathbf{W}^\top \mathbf{x}_t. & (3.13) \end{cases}$$

*In that situation, we will denote by $U^{p,q}_{\mathbf{W}} : I^{\mathbb{Z}} \longrightarrow I^{\mathbb{Z}}$ and $H^{p,q}_{\mathbf{W}} : I^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ 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 3.6** *Let $p(z) \in \mathbb{M}_{N,N}[z]$ be the polynomial given by*

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

*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 3.5 was proved using condition **(iii)** in the previous lemma then, any of the three conditions in that statement imply the echo state property for (3.12)-(3.13) 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 3.7** *Consider a non-homogeneous state-affine system as in (3.9)-(3.10) 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 3.6 then the corresponding reservoir filter has the fading memory property. More specifically, if* $p$ *satisfies condition* **(i)** *in Lemma 3.6, then* $H^{p,q}_{\mathbf{W}} : (I^{\mathbb{Z}_-}, \|\cdot\|_{w^\rho}) \longrightarrow \mathbb{R}$ *is continuous with* $w^\rho_t := (n_1 + 1)^{\rho t} \lambda^{\rho t}$ *and* $\rho \in (0, 1)$ *arbitrary. The same conclusion holds for conditions* **(ii)** *and* **(iii)** *with* $w^\rho_t := B^{\rho t}_p$ *and* $w^\rho_t := M^{\rho t}_p$*, 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

$$p_1(z) := A^1_0 + zA^1_1 + z^2 A^1_2 + \cdots + z^{n_1} A^1_{n_1}, \tag{3.14}$$

$$p_2(z) := A^2_0 + zA^2_1 + z^2 A^2_2 + \cdots + z^{n_2} A^2_{n_2}, \tag{3.15}$$

with $n_1, n_2 \in \mathbb{N}$, their direct sum $p_1 \oplus p_2(z) \in \mathbb{M}_{N_1+N_2, M_1+M_2}[z]$ and their Kronecker product $p_1 \otimes p_2(z) \in \mathbb{M}_{N_1 \cdot N_2, M_1 \cdot M_2}[z]$ is written as

$$p_1 \oplus p_2(z) = A^1_0 \oplus A^2_0 + zA^1_1 \oplus A^2_1 + z^2 A^1_2 \oplus A^2_2 + \cdots + z^{n_2} A^1_{n_2} \oplus A^2_{n_2} + z^{n_2+1} A^1_{n_2+1} \oplus 0 + \cdots + z^{n_1} A^1_{n_1} \oplus 0, \tag{3.16}$$

where we assumed that $n_2 \leq n_1$. Analogously,

$$p_1 \otimes p_2(z) = \sum_{i=0}^{n_1} \sum_{j=0}^{n_2} z^{i+j} A^1_i \otimes A^2_j. \tag{3.17}$$

**Proposition 3.8** *Let* $D \subset \mathbb{R}$ *be an open set and let* $H^{p_1,q_1}_{\mathbf{W}_1}, H^{p_2,q_2}_{\mathbf{W}_2} : D^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ *be two SAS reservoir functionals associated to two corresponding time-invariant SAS reservoir systems that have* $V_1$ *and* $V_2$ *as state spaces, respectively. Assume that the two polynomials with matrix coefficients* $p_1(z) \in \mathbb{M}_{N_1, M_1}[z]$ *and* $p_2(z) \in \mathbb{M}_{N_2, M_2}[z]$ *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* $1 > \epsilon > 0$*. Then, with the notation introduced in the expressions (3.16) and (3.17), we have that:*

**(i)** *For any* $\lambda \in \mathbb{R}$*, the linear combination* $H^{p_1,q_1}_{\mathbf{W}_1} + \lambda H^{p_2,q_2}_{\mathbf{W}_2}$ *is a SAS reservoir functional associated to a SAS that has* $V_1 \oplus V_2$ *as state space and:*

$$H^{p_1,q_1}_{\mathbf{W}_1} + \lambda H^{p_2,q_2}_{\mathbf{W}_2} = H^{p_1 \oplus p_2, q_1 \oplus q_2}_{\mathbf{W}_1 \oplus \lambda \mathbf{W}_2}. \tag{3.18}$$

**(ii)** *The product* $H^{p_1,q_1}_{\mathbf{W}_1}\cdot H^{p_2,q_2}_{\mathbf{W}_2}$ *is a SAS reservoir functional associated to a SAS that has* $V_1\oplus V_2\oplus(V_1\otimes V_2)$ *as state space and:*

$$H^{p_1,q_1}_{\mathbf{W}_1}\cdot H^{p_2,q_2}_{\mathbf{W}_2}=H^{p,q_1\oplus q_2\oplus(q_1\otimes q_2)}_{\mathbf{0}\oplus\mathbf{0}\oplus(\mathbf{W}_1\otimes\lambda\mathbf{W}_2)}, \tag{3.19}$$

*where* $p(z)\in\mathbb{M}_{N_{12}}[z]$, $N_{12}:=\dim V_1+\dim V_2+\dim V_1\cdot\dim V_2$, *is the polynomial with matrix coefficients whose block-matrix expression for the three summands in* $V_1\oplus V_2\oplus(V_1\otimes V_2)$ *is:*

$$p:=\begin{pmatrix} p_1 & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & p_2 & \mathbf{0}\\ p_1\otimes q_2 & q_1\otimes p_2 & p_1\otimes p_2\end{pmatrix}. \tag{3.20}$$

*The expression* $p_1\otimes q_2$ *(respectively,* $q_1\otimes p_2$*) denotes the linear map from* $V_1$ *(respectively,* $V_2$*) to* $V_1\otimes V_2$ *that associates to any* $v_1\in V_1$ *the element* $(p_1(z)v_1)\otimes q_2(z)$ *(respectively,* $q_1(z)\otimes(p_2(z)v_2)$*).*

*The equalities* (3.18) *and* (3.19) *show that the SAS family forms a polynomial algebra.*

**Theorem 3.9 (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* (3.9)-(3.10) *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* (3.5)*) 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 separate 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* $(I^{\mathbb{Z}_-},\|\cdot\|_{w^\rho})$ *which 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* $p(z)\in\mathbb{M}_{N,N}[z],q(z)\in\mathbb{M}_{N,1}[z]$ *with* $M_p,M_q<1-\epsilon$*, and a vector* $\mathbf{W}\in\mathbb{R}^N$ *such that*

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

**Remark 3.10** As it is stated in Theorem 3.9, it is the condition **(iii)** in Lemma 3.6 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 3.1, it is the algebras generated by them and not the families themselves that are universal.

**Remark 3.11** 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 85]. In discrete time, but only when the number of time steps is finite, this universal approximation property is exhibited [Flie 80] by homogeneous state-affine systems, that is, by setting $q(z)=0$ in (3.9)-(3.10).

**Remark 3.12 Generalization to multidimensional signals.** When the input signal is defined in $I^{\mathbb{Z}}_n$, 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 3.4, by polynomials of the form:

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

**Remark 3.13 SAS with trigonometric polynomials.** An analogous construction can be carried out using trigonometric polynomials of the type:

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

In this case, it is easy to establish that the resulting SAS family forms a polynomial algebra by invoking Proposition 3.8 and by reformulating the expressions (3.16) and (3.17) 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 3.9 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 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 see that the relation is true ***almost surely*** and we will abbreviate it ***a.s.*** We will denote by $L^\infty(\Omega,\mathbb{R}^n)$ the set of $\mathbb{R}^n$-valued random variables whose Euclidean norms have a finite essential supremum or that, equivalently, have almost surely bounded Euclidean norms. More specifically, let $\mathbf{X}:\Omega\longrightarrow\mathbb{R}^n$ be a random variable and let

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

We define

$$L^\infty(\Omega,\mathbb{R}^n):=\left\{\mathbf{X}:\Omega\longrightarrow\mathbb{R}^n\ \text{ random variable}\mid\|\mathbf{X}\|_{L^\infty}<\infty\right\}. \tag{4.2}$$

It can be shown that $L^\infty(\Omega,\mathbb{R}^n)$ is a Banach space (see [Ledo 91, pages 42 and 46], [Lord 14, page 149]. Given an element $\mathbf{X}\in L^\infty(\Omega,\mathbb{R}^n)$, 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 $\|\mathbf{X}\|^k$ of its norm are finite for all the elements $\mathbf{X}\in L^\infty(\Omega,\mathbb{R}^n)$.

**Lemma 4.1** *Let $\mathbf{X}\in L^\infty(\Omega,\mathbb{R}^n)$ and let $C>0$ be a real number. Then:*

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

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

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

**(iv)** *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 $L^\infty(\Omega, \mathbb{R}^n)$ as ***almost surely bounded*** random variables.

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

$$\begin{array}{rccc} \mathbf{z}: & \mathbb{Z}\times\Omega & \longrightarrow & \mathbb{R}^n \\ & (t,\omega) & \longmapsto & \mathbf{z}_t(\omega), \end{array} \tag{4.3}$$

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 $\mathbf{z}$. Additionally, we use the symbol $\mathcal{F}(\mathbf{z}) := \{\mathcal{F}_t(\mathbf{z}_t) \mid t \in \mathbb{Z}\}$, with $\mathcal{F}_t(\mathbf{z}_t) := \sigma(\{\mathbf{z}_s \mid s \leq t\})$ the sigma algebra generated by the process $\mathbf{z}$ up to time $t$, to refer to the ***filtration generated*** by $\mathbf{z}$. 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 (4.3) 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 [Come 06, Chapter 1]. Then, using the same prescription as in (4.1) with the supremum norm $\|\cdot\|_\infty$ in $(\mathbb{R}^n)^{\mathbb{Z}}$, we can define a norm $\|\cdot\|_{L^\infty}$ in the space of $\mathbb{R}^n$-valued stochastic processes by setting

$$\|\mathbf{z}\|_{L^\infty} := \operatorname*{ess\,sup}_{\omega\in\Omega} \|\mathbf{z}(\omega)\|_\infty = \operatorname*{ess\,sup}_{\omega\in\Omega} \left\{ \sup_{t\in\mathbb{Z}} \{\|\mathbf{z}_t(\omega)\|\} \right\}. \tag{4.4}$$

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 intertwined. The last statement contained in it complements part **(ii)** of Lemma 4.1 for processes.

**Lemma 4.2** *Let $\mathbf{z} : \Omega \longrightarrow (\mathbb{R}^n)^{\mathbb{Z}}$ be a stochastic process. Then,*

$$\|\mathbf{z}\|_{L^\infty} := \operatorname*{ess\,sup}_{\omega\in\Omega} \left\{ \sup_{t\in\mathbb{Z}} \{\|\mathbf{z}_t(\omega)\|\} \right\} = \sup_{t\in\mathbb{Z}} \left\{ \operatorname*{ess\,sup}_{\omega\in\Omega} \{\|\mathbf{z}_t(\omega)\|\} \right\}. \tag{4.5}$$

*Equivalently, using the notation in Lemma 4.1,*

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

*These equalities imply that for any positive real number $C > 0$, 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 $\sup_{t\in\mathbb{Z}} \|\mathbf{z}_t\|_{L^\infty} \leq C$.*

We now consider the space $L^\infty\left(\Omega, (\mathbb{R}^n)^{\mathbb{Z}}\right)$ of processes with finite $\|\cdot\|_{L^\infty}$ norm, that is,

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

We refer to the elements of $L^\infty\left(\Omega, (\mathbb{R}^n)^{\mathbb{Z}}\right)$ as ***almost surely bounded time series***. The following result shows that $L^\infty\left(\Omega, (\mathbb{R}^n)^{\mathbb{Z}}\right)$ coincides with the space of processes that take values in the Banach space $(\ell^\infty(\mathbb{R}^n), \|\cdot\|_\infty)$ and it is hence a Banach space.

**Lemma 4.3** *In the setup that we just introduced*

$$L^\infty\left(\Omega, (\mathbb{R}^n)^{\mathbb{Z}}\right) = L^\infty\left(\Omega, \ell^\infty(\mathbb{R}^n)\right) \tag{4.7}$$

*and $\left(L^\infty\left(\Omega, (\mathbb{R}^n)^{\mathbb{Z}}\right), \|\cdot\|_{L^\infty}\right)$ is a Banach space.*

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 (4.4) the $\ell^\infty$ 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:

$$\|\mathbf{z}\|_{L^\infty_w}:=\operatorname*{ess\,sup}_{\omega\in\Omega}\|\mathbf{z}(\omega)\|_w=\operatorname*{ess\,sup}_{\omega\in\Omega}\left\{\sup_{t\in\mathbb{Z}}\{\|\mathbf{z}_t(\omega)\|w_{-t}\}\right\}. \tag{4.8}$$

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. Since the space $\ell^\infty_w(\mathbb{R}^n)$ introduced in (2.2) is a Banach space [Grig 18] and a result similar to Lemma 4.3 shows that $L^\infty_w\left(\Omega,(\mathbb{R}^n)^{\mathbb{Z}_-}\right)=L^\infty\left(\Omega,\ell^\infty_w(\mathbb{R}^n)\right)$, we can conclude that $L^\infty_w\left(\Omega,(\mathbb{R}^n)^{\mathbb{Z}_-}\right)$ is a Banach space. Additionally, as in Lemma 4.2, we have that for any $\mathbf{z}\in L^\infty_w\left(\Omega,(\mathbb{R}^n)^{\mathbb{Z}_-}\right)$:

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

**Deterministic filters in a stochastic setup.** As we already pointed out, we consider filters $U$ that have almost surely bounded time series inputs and outputs, that is, if the inputs can take values in $\mathbb{R}^n$ then $U:L^\infty(\Omega,(\mathbb{R}^n)^{\mathbb{Z}})\longrightarrow L^\infty(\Omega,\mathbb{R}^{\mathbb{Z}})$. The same conventions as in the deterministic setup will be used in the identification of the different signals, namely, $\mathbf{z}$ will denote the filter inputs, the symbol $y$ is reserved for the output, and $d$ for the teaching signals. We emphasize that unlike in the deterministic case, the components of all these signals are this time random variables and not points in a Euclidean space.

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. The same applies to the correspondence between time-invariant filters $U:L^\infty(\Omega,(\mathbb{R}^n)^{\mathbb{Z}})\longrightarrow L^\infty(\Omega,\mathbb{R}^{\mathbb{Z}})$ and functionals $H_U:L^\infty(\Omega,(\mathbb{R}^n)^{\mathbb{Z}_-})\longrightarrow L^\infty(\Omega,\mathbb{R})$. We draw attention to the fact that assuming that the filters map into almost surely bounded processes is a genuine hypothesis that will need to be verified in each specific case considered.

We underline again that we will restrict our attention to intrinsically ***deterministic filters*** that are obtained by presenting almost surely bounded stochastic inputs $\mathbf{z}\in L^\infty(\Omega,(\mathbb{R}^n)^{\mathbb{Z}_-})$ to causal and time-invariant filters $H:(\mathbb{R}^n)^{\mathbb{Z}_-}\longrightarrow\mathbb{R}$ similar to those introduced in the previous section which, in some cases, yields maps of the type $H:L^\infty(\Omega,(\mathbb{R}^n)^{\mathbb{Z}_-})\longrightarrow L^\infty(\Omega,\mathbb{R})$ and that we will denote with the same symbol. In that situation, the dependence on the probability space of the image $(H(\mathbf{z}))(\omega)$ takes place exclusively through the dependence $\mathbf{z}(\omega)$ in the input. Causal and time-invariant deterministic filters with stochastic inputs produce almost surely causal and time-invariant filters. Moreover, for any $t\in\mathbb{Z}$, the random variable $U_H(\mathbf{z})_t:=H(\mathbb{P}_{\mathbb{Z}_-}\circ U_t(\mathbf{z}))=H\left(\mathbf{z}_t,\mathbf{z}_{t-1},\mathbf{z}_{t-2},\ldots\right)$ is clearly $\mathcal{F}_t(\mathbf{z}_t)$-measurable.

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

Let $M>0$ and define, using Lemma 4.2,

$$K^{L^\infty}_M:=\left\{\mathbf{z}\in L^\infty(\Omega,(\mathbb{R}^n)^{\mathbb{Z}_-})\mid\|\mathbf{z}\|_{L^\infty}\leq M\right\}=\left\{\mathbf{z}\in L^\infty(\Omega,(\mathbb{R}^n)^{\mathbb{Z}_-})\mid\|\mathbf{z}_t\|_{L^\infty}\leq M,\ \text{for all } t\in\mathbb{Z}_-\right\}. \tag{4.10}$$

The sets $K^{L^\infty}_M$ are the stochastic counterparts of the sets $K_M$ in the deterministic setup; we will say that $K^{L^\infty}_M$ is a set of ***almost surely uniformly bounded processes***. A stochastic analog of Lemma 2.1 can be formulated for them with $K_M$ replaced by $K^{L^\infty}_M$, the norm $\|\cdot\|$ by $\|\cdot\|_{L^\infty}$, and the weighted norm $\|\cdot\|_w$ by $\|\cdot\|_{L^\infty}$.

The following result shows that the fading memory and the universality properties are naturally inherited by deterministic filters with almost surely bounded inputs.

**Theorem 4.4** *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 (4.10), respectively. The following properties hold true:*

**(i)** *Let $H : (K_M, \|\cdot\|_w) \longrightarrow \mathbb{R}$ be a causal and time-invariant filter. Then $H : (K_M, \|\cdot\|_w) \longrightarrow \mathbb{R}$ 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 intputs 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 result is a stochastic analog of Theorem 3.1 and 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. It is easy to verify that if a reservoir system with uniformly bounded deterministic inputs satisfies the echo state property and the FMP then so does its counterpart with almost surely uniformly bounded inputs.

**Theorem 4.5** *Let $M > 0$ and let $K_M^{L^\infty}$ be the set of almost surely uniformly bounded processes introduced in (4.10). Consider the set $\mathcal{R}$*

$$\mathcal{R} := \{H_{h_i}^{F_i} : K_M^{L^\infty} \longrightarrow L^\infty(\Omega, \mathbb{R}) \mid h_i \in \mathrm{Pol}(\mathbb{R}^{N_i}, \mathbb{R}), F_i : \mathbb{R}^{N_i} \times \mathbb{R}^n \longrightarrow \mathbb{R}^{N_i}, i \in I, N_i \in \mathbb{N}\} \tag{4.11}$$

*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 contains all the constant functionals, then any deterministic, causal, time-invariant fading memory filter $H : (K_M^{L^\infty}, \|\cdot\|_{L^\infty_w}) \longrightarrow L^\infty(\Omega, \mathbb{R})$ can be uniformly approximated by elements in $\mathcal{A}(\mathcal{R})$, that is, for any $\epsilon > 0$, there exist indices $\{i_1, \ldots, i_r\} \subset I$ and a polynomial $p : \mathbb{R}^r \longrightarrow \mathbb{R}$ such that*

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

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 4.5 is the stochastic analog of Corollary 3.2.

**Corollary 4.6** *Let $M > 0$ and let $K_M^{L^\infty}$ be the set of almost surely uniformly bounded processes introduced in (4.10). Let*

$$\mathcal{R}_w := \{H_h^F : K_M^{L^\infty} \longrightarrow L^\infty(\Omega, \mathbb{R}) \mid h \in \mathrm{Pol}(\mathbb{R}^N, \mathbb{R}), F : \mathbb{R}^N \times \mathbb{R}^n \longrightarrow \mathbb{R}^N, N \in \mathbb{N}\} \tag{4.12}$$

*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 $\|\cdot\|_{L^\infty_w}$. Then $\mathcal{R}_w$ is universal, that is, for any time-invariant fading memory filter $H : (K_M^{L^\infty}, \|\cdot\|_{L^\infty_w}) \longrightarrow L^\infty(\Omega, \mathbb{R})$ 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 (3.6)-(3.7) suffices to achieve universality. The proof of the following statement is a direct consequence of Corollary 3.3 and Theorem 4.4.

**Corollary 4.7** *Let $M > 0$ and let $K_M^{L^\infty}$ be the set of almost surely uniformly bounded processes introduced in (4.10). Let $\mathcal{L}_\epsilon$ be the family introduced in Corollary 3.3 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 $H : (K_M^{L^\infty}, \|\cdot\|_{L^\infty_{w^\rho}}) \longrightarrow L^\infty(\Omega,\mathbb{R})$ and any $\varepsilon > 0$, there exists $H_p^{A,\mathbf{c}} \in \mathcal{L}_\epsilon$ such that $\|H - H_p^{A,\mathbf{c}}\|_\infty := \sup_{\mathbf{z}\in K_M^{L^\infty}} \|H(\mathbf{z}) - H_p^{A,\mathbf{c}}(\mathbf{z})\|_{L^\infty} < \varepsilon$.*

**Remark 4.8** 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 4.7 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 strong vector autoregressive model (VAR) of order 1. This statement applies, for example, to any stationary ARMA [Box 76, Broc 06] or GARCH [Engl 82, Boll 86, Fran 10] model 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 3.9 and the equivalence stated in Theorem 4.4.

**Theorem 4.9 (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,*

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

*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 (3.9)-(3.10) and that satisfy $M_p := \max_{z\in I} \|p(z)\| < 1-\epsilon$ and $M_q := \max_{z\in I} \|q(z)\| < 1-\epsilon$, with $I := [-1,1]$. The family $\mathcal{S}_\epsilon$ forms a polynomial subalgebra of $\mathcal{R}_{w^\rho}$ (as defined in (4.12)) 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*

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

# 5 Echo state affine systems (ESAS) and their forecasting performance

Echo state networks (ESNs) were introduced in [Jaeg 04] as one of the first families of reservoir computing architectures and have been very successful in many information processing tasks. An ESN is determined by the following state-space transformation:

$$\begin{cases} \mathbf{x}_t = \sigma\left(\boldsymbol{\alpha} + A\mathbf{x}_{t-1} + \gamma \mathbf{c}\mathbf{z}_t\right), & (5.1) \\ y_t = \mathbf{W}^\top \mathbf{x}_t. & (5.2) \end{cases}$$

In these equations $\mathbf{z} \in K_M \subset (\mathbf{R}^n)^{\mathbb{Z}}$ is the input signal, $\mathbf{c} \in \mathbb{M}_{N,n}$ is called the ***input mask***, $\boldsymbol{\alpha} \in \mathbb{R}^N$ is the ***input shift***, and $\gamma \in \mathbb{R}$ is the ***input gain***. The matrix $A \in \mathbb{M}_{N,N}$ is referred to as the ***reservoir matrix***. The map $\sigma$ in the state-space equation (5.1) is constructed by componentwise application of a sigmoid function (like the hyperbolic tangent or the logistic function) and is called the ***activation function***. It can be shown (see [Jaeg 10]) that (5.1)-(5.2) induces a causal, time-invariant, and fading memory filter whenever $\|A\|_2 = \sigma_{\max}(A) < 1$. Moreover, the fading memory property does not hold whenever the spectral radius $\rho(A)$ of $A$ exceeds one.

ESNs are implemented in practice by generating a large random reservoir matrix $A$ and an input mask $\mathbf{c}$ and by tuning the input shift, input gain, and the spectral radius of $A$ so that the performance in the execution of a specific task is maximized. Given specific input $\mathbf{z} \in K_M$ and teaching signals $\mathbf{d} \in \mathbb{R}^{\mathbb{Z}}$, the ***readout vector*** $\mathbf{W}$ is obtained by solving the ridge regularized regression problem:

$$\mathbf{W} = \underset{\widetilde{\mathbf{W}} \in \mathbb{R}^N}{\arg\min} \left( \lim_{t\to\infty} \frac{1}{2t} \sum_{i=-t}^{t} (\widetilde{\mathbf{W}}^\top \mathbf{x}_i - d_i)^2 + \lambda \|\mathbf{W}\|^2 \right),$$

where $\lambda > 0$ is the regularization constant.

The interesting universality properties associated to the non-homogeneous state-affine systems that we proved in Section 3.2 suggest a generalization of the ESNs that we denominate ***echo state affine systems (ESAS)*** and that are determined by the state-space equations:

$$\begin{cases} \mathbf{x}_t = \sigma\left(p(z_t)\mathbf{x}_{t-1} + q(z_t)\right), & (5.3) \\ y_t = \mathbf{W}^\top \mathbf{x}_t. & (5.4) \end{cases}$$

An ESAS is hence a SAS in which an activation function has been added following the pattern of the ESN case. ESAS are a strict generalization of ESNs: an ESAS for which the degree of the $p$ polynomial is zero and the degree of the $q$ polynomial is one, is an ESN.

**A forecasting exercise.** A task in which ESNs have been particularly successful is the forecasting of chaotic time series generated by the the Mackey-Glass system [Mack 77] characterized by the time-delay differential equation:

$$\frac{dx}{dt} = \frac{0.2\, x(t-\tau)}{1 + x(t-\tau)^{10} - 0.1\, x(t)}. \tag{5.5}$$

In our forecasting exercise we use for the delay $\tau$ the customary value in the forecasting literature of $\tau = 17$. The Lyapunov exponent of this system has been numerically estimated in [Jaeg 04], where values around $6 \cdot 10^{-3}$ were obtained, which shows that this system does indeed generate chaotic time series.

Forecasting of those time series has been conducted using ESNs as in [Jaeg 04] by using a sparse (1% connectivity) reservoir matrix $A$ with dimension $N = 1000$ and with entries randomly drawn from a

uniform distribution in the interval $[-1, 1]$. The training is carried out by using as input $\mathbf{z}$ the sampling with step size 1.0 of a solution of (5.5) with length 5000 and obtained from a differential equation solver that produced an absolute accuracy of $1 \cdot 10^{-16}$. The teaching signal is set to $\mathbf{d} = U_{-1}\mathbf{z}$, that is, the input signal shifted in time one step into the future. Forecasting is then carried out with the trained system running autonomously by feeding the output into the input. The trained ESN is used in that case as a proxy for the original system (5.5) that has been hence *learned* by (5.1)-(5.2). Using this approach, the results reported in [Jaeg 04] improved by a factor of 700 the performances previously documented in the literature.

We have checked if various ESAS architectures are able to improve the ESN performance in this task. In order to do so and to make the two systems comparable, we have used ESAS of the same dimension ($N = 1000$), the reservoir matrices in the $p$ polyomial and the ESN reservoir matrix $A$ are sparse and have the same connectivity (1%) and, more importantly, the same hyperparameters have been optimized. More specifically, in the ESN case we optimized the input shift and gain and we also tuned the spectral radius of $A$. The same procedure was carried out in the ESAS case where, additionally, we tuned the spectral radii of each of the reservoir matrices in the $p$ polyomial.

The results of these forecasting exercises are depicted in Figure 1 where we plot the normalized root mean square error (NRMSE) committed by the ESN and by several ESAS($r$,$s$) arquitectures ($r$ and $s$ stand for the degrees of the $p$ and $q$ polynomials, respectively) as a function of the forecasting horizon. The NRMSE has been averaged over 2000 points. This figure shows that the chosen ESAS architectures systematically outperform the ESN. A particularly significant improvement is exhibited by the ESAS(1,2) that, at a forecasting horizon of 121, reduces the NRMSE of 39.18%.

# 6 Appendices

## 6.1 Proof of Lemma 2.1

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:

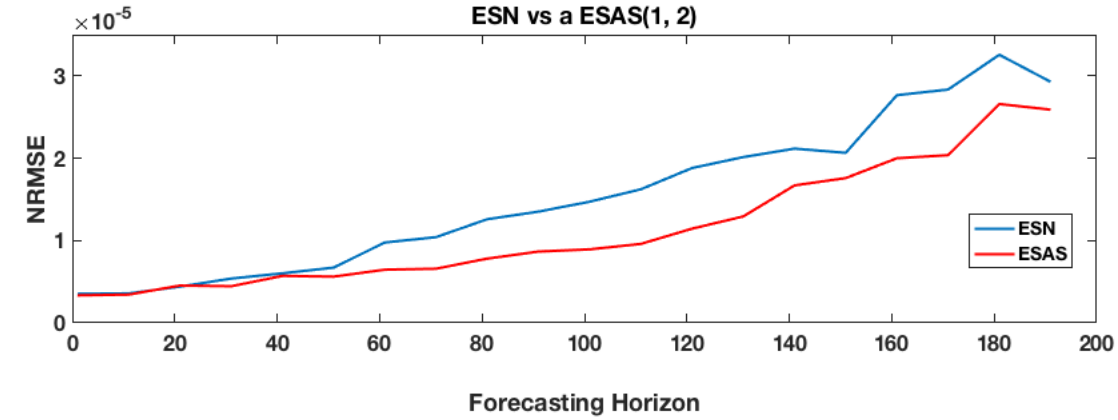
$$\begin{aligned}\sum_{t=0}^{\infty} \|\mathbf{z}_{-t}\| w_t = \sum_{t=0}^{\infty} \|\mathbf{z}_{-t}\| \lambda^t = \sum_{t=0}^{\infty} \|\mathbf{z}_{-t}\| (\lambda^{1-\rho}\lambda^{\rho})^t = \sum_{t=0}^{\infty} \|\mathbf{z}_{-t}\| \lambda^{(1-\rho)t}\lambda^{\rho t} \\ \leq \sum_{t=0}^{\infty} \sup_{i\in\mathbb{N}}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

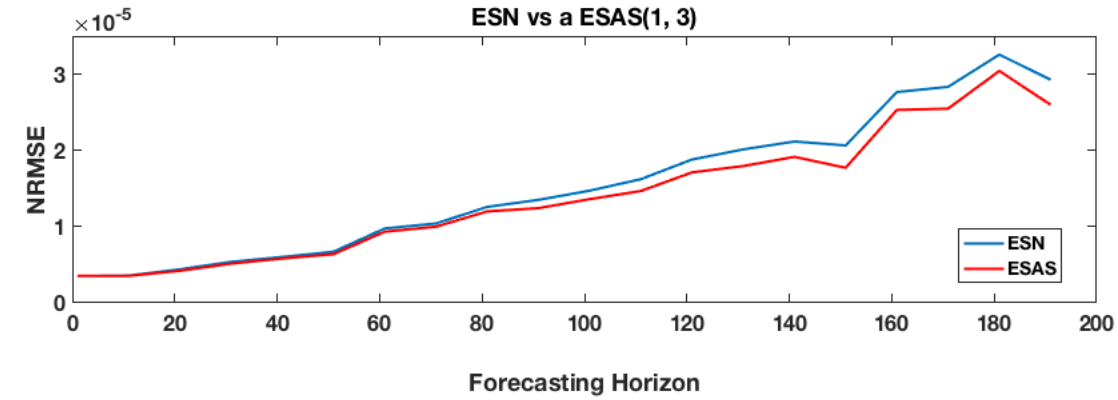
which proves (2.4). The proof of (2.5) is similar and follows from noticing that:

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

## 6.2 Proof of Lemma 2.2

We recall first that by Lemma 2.1 we have that $\|\mathbf{z}\|_w < \infty$, for any $\mathbf{z} \in K_M$. Second, since $(D^{\mathbb{Z}_-}, \|\cdot\|_w)$ is a normed space it is hence metrizable and therefore so is $(K_M, \|\cdot\|_w)$ when endowed with the relative topology (see, for instance, [Munk 14, Exercise 1, page 131]). We will then conclude the compactness



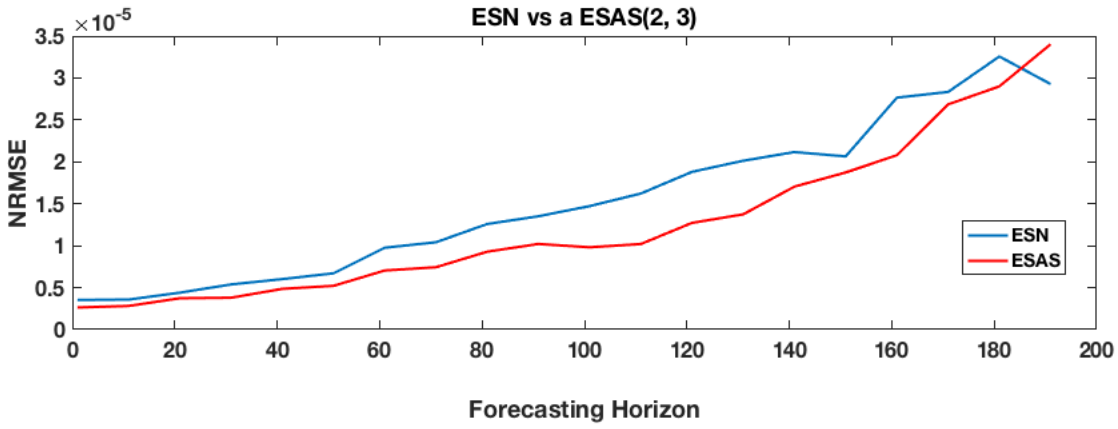




Figure 1: Results of the Mackey-Glass chaotic time series forecasting exercise. The different panels depict the normalized root mean square error (NRMSE) committed by the ESN and by several ESAS($r$,$s$) arquitectures ($r$ and $s$ stand for the degrees of the $p$ and $q$ polynomials, respectively) as a function of the forecasting horizon. The NRMSE has been averaged over 2000 points. The chosen ESAS architectures systematically outperform the ESN. A particularly significant improvement is exhibited by the ESAS(1,2) that at a forecasting horizon of 121 reduces the NRMSE of 39.18%.

of $(K_M, \|\cdot\|_w)$ by showing that this space is sequentially compact (see, for example [Munk 14, Theorem 28.2]). We proceed by using the strategy in the proof of Lemma 1 in [Boyd 85].

For any $n \in \mathbb{N}$, let $K_M^n$ be the set obtained by projecting into $D^{\{-n,\ldots,-1,0\}}$ the elements of $K_M \subset D^{\mathbb{Z}_-}$. Given an element $\mathbf{z} \in K_M$, we will denote by $\mathbf{z}^{(n)} := (\mathbf{z}_{-n}, \ldots, \mathbf{z}_0)$ its projection into $K_M^n$. Additionally, notice that $K_M^n = [-M, M]^n$ is a compact (and hence sequentially compact) topological space with the product topology.

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 and 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 2.6

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

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

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 3.1

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 2.2, the set $(K_M, \|\cdot\|_w)$ is compact, the Stone-Weierstrass theorem (Theorem 7.3.1 in [Dieu 69]) implies that $\mathcal{A}(\mathcal{R})$ is dense in $(C^0(K_M), \|\cdot\|_w)$, which concludes the proof. ■

## 6.5 Proof of Corollary 3.3

In order to show that the reservoir systems in $\mathcal{L}_\epsilon$ induce reservoir filters, we will show that they have the echo state property by using the following lemma, whose proof can be found in [Grig 18].

**Lemma 6.1** *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:*

$$\|F(\mathbf{x}, \mathbf{z}) - F(\mathbf{y}, \mathbf{z})\| \leq r \|\mathbf{x} - \mathbf{y}\|, \quad \textit{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^N \textit{ and all } \mathbf{z} \in \mathbb{R}^n,$$

*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 $F(\mathbf{x}, \mathbf{z}) := A\mathbf{x} + \mathbf{cz}$ associated to (3.6) is a contracting map with constant $\sigma_{\max}(A)$ which, by hypothesis, is smaller than one. Indeed,

$$\|F(\mathbf{x}, \mathbf{z}) - F(\mathbf{y}, \mathbf{z})\| = \|A(\mathbf{x} - \mathbf{y})\| \leq \sigma_{\max}(A) \|\mathbf{x} - \mathbf{y}\| \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^N \text{ and all } \mathbf{z} \in \mathbb{R}^n.$$

By Lemma 6.1 we can conclude that this reservoir system has a reservoir filter associated that we now show is explicitly given by (3.8). 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 (3.8) is convergent. Let $n, m \in \mathbb{N}$ 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 (3.8) is a solution of the recursions (3.6)-(3.7) 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

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

and that the time series $\widetilde{x_t}$ defined by $\widetilde{x_t} := \sum_{i=0}^{\infty} A^i \mathbf{c} z_{t-i}$ satisfies the recursion relation (3.6).

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

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

At the same time,

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

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 $\lambda_\rho$-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 $h:\mathbb{R}^N\to\mathbb{R}$ implies that for any $\varepsilon>0$ there exists an element $\delta(\varepsilon)>0$ such that for any $\mathbf{v}\in\mathbb{R}^N$ 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,

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

If we now use (2.5) in Lemma 2.1 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 $\lambda_\rho$-exponential fading memory filter.

In order to establish the universality statement in the corollary we will proceed, as in the proof of Theorem 3.1, 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 $(C^0(K_M),\|\cdot\|_{w^\rho})$.

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 $\sigma_{\max}(A_1),\sigma_{\max}(A_1)<1-\epsilon$, 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 $\mathbf{c}_i \in \mathbb{M}_{N_i,n_i}$, $i \in \{1,2\}$ 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 (3.2)-(3.3) 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 $K_M$. 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 $U^{A,\mathbf{c}} : (\mathbb{R}^n)^{\mathbb{Z}_-} \longrightarrow (\mathbb{R}^n)^{\mathbb{Z}_-}$ be the linear filter associated to $A$ and $\mathbf{c}$ via the recursion (3.6). Using the preceding arguments we have that

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

At the same time, it is easy to verify that the filter $U^{\mathbb{I}_n - A,\mathbf{c}} : (\mathbb{R}^n)^{\mathbb{Z}_-} \longrightarrow (\mathbb{R}^n)^{\mathbb{Z}_-}$ given by $\left(U^{\mathbb{I}_n - A,\mathbf{c}}(\mathbf{z})\right)_t = \mathbf{z}_t - A\mathbf{z}_{t-1}$, is the two-sided inverse of $U^{A,\mathbf{c}}$, that is, $U^{A,\mathbf{c}} \circ U^{\mathbb{I}_n - A,\mathbf{c}} = U^{\mathbb{I}_n - A,\mathbf{c}} \circ U^{A,\mathbf{c}} = \mathbb{I}_{(\mathbb{R}^n)^{\mathbb{Z}_-}}$, which guarantees the bijectivity of $U^{A,\mathbf{c}}$ and implies in passing that $U^{A,\mathbf{c}}(\mathbf{z}_1) \neq U^{A,\mathbf{c}}(\mathbf{z}_2)$. Since polynomials in $\mathbb{R}^n$ separate points, there exists a polynomial readout map $h : \mathbb{R}^n \longrightarrow \mathbb{R}$ such that $h\left(U^{A,\mathbf{c}}(\mathbf{z}_1)\right) \neq h\left(U^{A,\mathbf{c}}(\mathbf{z}_2)\right)$ which shows that the associated functional $H_h^{A,\mathbf{c}}$ is such that

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

## 6.6 Proof of Proposition 3.5

We start by noting, as we did in the proof of Corollary 3.3, that the condition (3.11) implies that the reservoir map associated to (3.9) is a contraction and hence, by Lemma 6.1, it satisfies the echo state property and has a well-defined associated filter.

We now prove that the condition (3.11) implies the convergence of the series in the expression (3.12). 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 $n, m \in \mathbb{N}$ 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 (3.12) 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 3.3. We conclude by proving

that (3.12) is indeed a solution of (3.9):

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

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

## 6.7 Proof of Lemma 3.6

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

## 6.8 Proof of Proposition 3.7

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

**Lemma 6.2** *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. Then, for any $\mathbf{x},\mathbf{y}\in U$ we have:*

$$\|\mathbf{f}(\mathbf{y})-\mathbf{f}(\mathbf{x})\|_2\leq\sqrt{nm}\max_{i\in\{1,\ldots,n\}}\left(\sup_{\mathbf{z}\in U}\|\partial_i\mathbf{f}(\mathbf{z})\|_2\right)\|\mathbf{x}-\mathbf{y}\|. \tag{6.9}$$

**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 13]) that

$$\|A\|_2\leq\|A\|_F\leq\sqrt{r}\,\|A\|_2, \tag{6.10}$$

where $r$ is the rank of $A$. Consider now $\mathbf{x},\mathbf{y}\in U$ arbitrary and let $D\mathbf{f}(\mathbf{z}):R^n\longrightarrow\mathbb{M}_m$ 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 [Abra 88]) and hence:

$$\|\mathbf{f}(\mathbf{y})-\mathbf{f}(\mathbf{x})\|_F\leq\sup_{t\in[0,1]}\|D\mathbf{f}((1-t)\mathbf{x}+t\mathbf{y})\|_2\,\|\mathbf{x}-\mathbf{y}\|. \tag{6.11}$$

The first inequality in (6.10) and (6.11) imply that

$$\|\mathbf{f}(\mathbf{y})-\mathbf{f}(\mathbf{x})\|_2\leq\sup_{\mathbf{z}\in U}\|D\mathbf{f}(\mathbf{z})\|_2\,\|\mathbf{x}-\mathbf{y}\|. \tag{6.12}$$

At the same time, notice that by (6.10)

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

This inequality together with (6.12) imply the statement (6.9) 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 3.6 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:

$$\begin{aligned}|H^{p,q}_{\mathbf{W}}(\mathbf{z}) - H^{p,q}_{\mathbf{W}}(\mathbf{s})| &= \left| \mathbf{W}^\top \left[ \sum_{j=0}^{\infty} \left( \left( \prod_{k=0}^{j-1} p(z_{-k}) \right) q(z_{-j}) - \left( \prod_{k=0}^{j-1} p(s_{-k}) \right) q(s_{-j}) \right) \right] \right| \\ &\leq \|\mathbf{W}\| \sum_{j=0}^{\infty} \left\| a_j(\underline{z_{-j+1}}) q(z_{-j}) - a_j(\underline{s_{-j+1}}) q(s_{-j}) \right\|, \quad \text{where} \quad a_j(\underline{z_{-j+1}}) := \prod_{k=0}^{j-1} p(z_{-k}). \end{aligned} \tag{6.13}$$

We now bound the right hand side of (6.13) as follows:

$$\begin{aligned} &\sum_{j=0}^{\infty} \left\| a_j(\underline{z_{-j+1}}) q(z_{-j}) - a_j(\underline{s_{-j+1}}) q(s_{-j}) \right\| \\ &\quad = \sum_{j=0}^{\infty} \left\| a_j(\underline{z_{-j+1}}) q(z_{-j}) + a_j(\underline{z_{-j+1}}) q(s_{-j}) - a_j(\underline{z_{-j+1}}) q(s_{-j}) - a_j(\underline{s_{-j+1}}) q(s_{-j}) \right\| \\ &\quad \leq \sum_{j=0}^{\infty} \left\| a_j(\underline{z_{-j+1}}) \right\|_2 \|q(z_{-j}) - q(s_{-j})\| + \left\| a_j(\underline{z_{-j+1}}) - a_j(\underline{s_{-j+1}}) \right\|_2 \|q(s_{-j})\| \end{aligned} \tag{6.14}$$

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

$$\left\| a_j(\underline{z_{-j+1}}) \right\|_2 \|q(z_{-j}) - q(s_{-j})\| \leq M_p^j L_q \left| z_{-j} - s_{-j} \right|, \tag{6.15}$$

which inserted in (6.14) and in (6.13) implies that

$$|H^{p,q}_{\mathbf{W}}(\mathbf{z}) - H^{p,q}_{\mathbf{W}}(\mathbf{s})| \leq \|\mathbf{W}\| L_q \left[ \sum_{j=0}^{\infty} M_p^j \left| z_{-j} - s_{-j} \right| + \sum_{j=0}^{\infty} \left\| a_j(\underline{z_{-j+1}}) - a_j(\underline{s_{-j+1}}) \right\|_2 \right] \tag{6.16}$$

We now bound above the second summand in (6.16) using the inequality (6.9) in the statement of Lemma 6.2 as well as the following straightforward identity:

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

Using this relation we can write:

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

where the last inequality is a consequence of (6.9). Let $M_{p'} := \sqrt{N}\sup_{z\in I}\left(\|p'(z)\|_2\right)$, 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 the Cauchy product of the series with general terms $M_p^n$ and $M_p^n\,|z_{-n} - s_{-n}|$, we can conclude using Merten's Theorem [Apos 74, Theorem 8.46] that

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

If we now substitute this relation in (6.16) and we use Lemma 2.1 with weighting sequences $w_t^\rho := M_p^{\rho t}$, for any $\rho \in (0,1)$, we obtain that:

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

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 3.8

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 (3.2)-(3.3), 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 (3.10):

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

Using (3.9) 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

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

which proves (3.20) and hence (3.19).

In order to show that the reservoir functionals on the right hand side of (3.18) and (3.19) are well-defined we prove the following lemma.

**Lemma 6.3** *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 $1 > \epsilon > 0$. 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.** Suppose that the polynomials $p_1$ and $p_2$ are defined on the vector spaces $V_1$ and $V_2$, respectively. Let $\mathbf{x} = \mathbf{x}_1 \oplus \mathbf{x}_2 \in V_1 \oplus V_2$, with $\mathbf{x}_1 \in V_1$ and $\mathbf{x}_2 \in V_2$. 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

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

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 3.5 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 $H^{p, q_1 \oplus q_2 \oplus (q_1 \otimes q_2)}_{\mathbf{0} \oplus \mathbf{0} \oplus (\mathbf{W}_1 \otimes \lambda \mathbf{W}_2)}$ because due to the block diagonal character of (3.20) 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 6.3 we can conclude that $\|p(z)\|_2 < 1 - \epsilon$ for all $z \in [-1, 1]$ and, again by Proposition 3.5, the reservoir functional $H^{p, q_1 \oplus q_2 \oplus (q_1 \otimes q_2)}_{\mathbf{0} \oplus \mathbf{0} \oplus (\mathbf{W}_1 \otimes \lambda \mathbf{W}_2)}$ is well-defined. ■

## 6.10 Proof of Theorem 3.9

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 3.5 and 3.7, that this family is made of time-invariant reservoir filters that have the FMP with respect to weighting sequences of the form $w_t^p := M_p^{\rho t}$, $\rho \in (0, 1)$. Additionally, using Lemma 2.6 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 $w_t^\rho := (1-\epsilon)^{\rho t}$, $\rho \in (0, 1)$.

We now show that the elements in $\mathcal{S}_\epsilon$ form a polynomial algebra as a consequence mainly of Lemma 6.3 and Proposition 3.8. Indeed, if $H^{p_1, q_1}_{\mathbf{W}_1}, H^{p_2, q_2}_{\mathbf{W}_2} : I^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ are two elements in $\mathcal{S}_\epsilon$ then by (3.18) and part **(i)** of Lemma 6.3 so is $H^{p_1, q_1}_{\mathbf{W}_1} + \lambda H^{p_2, q_2}_{\mathbf{W}_2}$, for any $\lambda \in \mathbb{R}$. The same conclusion holds for the product $H^{p_1, q_1}_{\mathbf{W}_1} \cdot H^{p_2, q_2}_{\mathbf{W}_2}$ because due to the block diagonal character of (3.20) 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 6.3 we can conclude that $\|p(z)\|_2 < 1 - \epsilon$ for all $z \in [-1, 1]$. The same fact can be stated about $q_1 \oplus q_2 \oplus (q_1 \otimes q_2)$ and hence by (3.19) we obtain that the product $H^{p_1, q_1}_{\mathbf{W}_1} \cdot H^{p_2, q_2}_{\mathbf{W}_2}$ belongs to $\mathcal{S}_\epsilon$.

Finally, we observe that 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 $\mathcal{L}_\epsilon$ and 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 $U^{A, \mathbf{c}}(\mathbf{z}_1) \neq U^{A, \mathbf{c}}(\mathbf{z}_2)$, this implies the existence of a $t \in \mathbb{Z}$ such that $(U^{A, \mathbf{c}}(\mathbf{z}_1))_t \neq (U^{A, \mathbf{c}}(\mathbf{z}_2))_t$. The point separation property follows from choosing any vector $\mathbf{W} \in \mathbb{R}^N$ such that $\mathbf{W}^\top (U^{A, \mathbf{c}}(\mathbf{z}_1))_t \neq \mathbf{W}^\top (U^{A, \mathbf{c}}(\mathbf{z}_2))_t$, which implies that $(U^{A, \mathbf{c}}_{\mathbf{W}}(\mathbf{z}_1))_t \neq (U^{A, \mathbf{c}}_{\mathbf{W}}(\mathbf{z}_2))_t$. Therefore, $U^{A, \mathbf{c}}_{\mathbf{W}}(\mathbf{z}_1) \neq U^{A, \mathbf{c}}_{\mathbf{W}}(\mathbf{z}_2)$ and $H_{U^{A, \mathbf{c}}_{\mathbf{W}}}(\mathbf{z}_1) \neq H_{U^{A, \mathbf{c}}_{\mathbf{W}}}(\mathbf{z}_2)$, 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 $H^{p, q}_{\mathbf{W}}(\mathbf{z}) = \mathbf{W}^\top q$, 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 2.2. ■

## 6.11 Proof of Lemma 4.1

**(i)** Let $A := \{\rho \in \mathbb{R}_+ \mid \|\mathbf{X}\| < \rho \quad \text{almost surely}\}$. It suffices to show that $\|\mathbf{X}\|_{L^\infty} := \inf A \in A$, which implies that $\|\mathbf{X}\| \leq \|\mathbf{X}\|_{L^\infty}$ almost surely. Indeed, consider the sequence $a_j := \|\mathbf{X}\|_{L^\infty} + 1/j$, $j \in \mathbb{N}$. 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 $F := \{\omega \in \Omega \mid \|\mathbf{X}(\omega)\| > \|\mathbf{X}\|_{L^\infty}\}$ and $F_j := \{\omega \in \Omega \mid \|\mathbf{X}(\omega)\| > \rho_j\}$. 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 [Grim 01, Lemma 5, page 7]) $\lim_{j\to\infty} \mathbb{P}(F_j) = \mathbb{P}(F)$. 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)**, $\|\mathbf{X}\| \leq \|\mathbf{X}\|_{L^\infty} \leq C$ almost surely. Conversely, if $\|\mathbf{X}\| \leq C$ almost surely, then $C \in A = \{\rho \in \mathbb{R}_+ \mid \|\mathbf{X}\| < \rho \quad \text{almost surely}\}$. Consequently, $\|\mathbf{X}\|_{L^\infty} = \inf A \leq C \in A$, as required.
**(iii)** Suppose first that $\|\mathbf{X}\| \leq C$ almost surely and define $F := \{\omega \in \Omega \mid \|\mathbf{X}(\omega)\| > C\}$. By hypothesis, we have that $\mathbb{P}(F) = 0$ and $\mathbb{P}(\Omega \setminus F) = 1$. Then,

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

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

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

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 [Grim 01, Lemma 5, page 7]) $\lim_{n\to\infty} \mathbb{P}(F_n) = \mathbb{P}(F)$. Now,

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

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 $\|\mathbf{X}\| \leq C$ a.s.
**(iv)** 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 4.2

Suppose first that

$$\operatorname*{ess\,sup}_{\omega\in\Omega} \left\{\sup_{t\in\mathbb{Z}} \{\|\mathbf{z}_t(\omega)\|\}\right\} < \sup_{t\in\mathbb{Z}} \left\{\operatorname*{ess\,sup}_{\omega\in\Omega} \{\|\mathbf{z}_t(\omega)\|\}\right\}. \tag{6.17}$$

By the approximation property of the supremum [Apos 74, Theorem 1.14], there exists $t_0 \in \mathbb{Z}$ such that

$$\operatorname*{ess\,sup}_{\omega\in\Omega} \left\{\sup_{t\in\mathbb{Z}} \{\|\mathbf{z}_t(\omega)\|\}\right\} < \operatorname*{ess\,sup}_{\omega\in\Omega} \{\|\mathbf{z}_{t_0}(\omega)\|\} \leq \sup_{t\in\mathbb{Z}} \left\{\operatorname*{ess\,sup}_{\omega\in\Omega} \{\|\mathbf{z}_t(\omega)\|\}\right\}. \tag{6.18}$$

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 4.1

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

Now, by part **(ii)** in Lemma 4.1, this implies that

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

However, this expression is in contradiction with the first inequality in (6.18) and hence the assumption (6.17) cannot be correct. This argument implies that

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

We now prove the reverse inequality, that is,

$$\operatorname*{ess\,sup}_{\omega\in\Omega} \left\{ \sup_{t\in\mathbb{Z}} \{\|\mathbf{z}_t(\omega)\|\} \right\} \leq \sup_{t\in\mathbb{Z}} \left\{ \operatorname*{ess\,sup}_{\omega\in\Omega} \{\|\mathbf{z}_t(\omega)\|\} \right\}. \tag{6.19}$$

By part **(ii)** of Lemma 4.1, this inequality holds if and only if

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

Now, by part **(i)** in Lemma 4.1, we have that $\|\mathbf{z}_t(\omega)\| \leq \operatorname{ess\,sup}_{\omega\in\Omega} \{\|\mathbf{z}_t(\omega)\|\}$, almost surely and for each fixed $t \in \mathbb{Z}$. Let $A_t \subset \Omega$ be the zero-measure set such that $\|\mathbf{z}_t(\omega)\| > \operatorname{ess\,sup}_{\omega\in\Omega} \{\|\mathbf{z}_t(\omega)\|\}$ 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

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

Since $B$ has measure one, this implies that

$$\sup_{t\in\mathbb{Z}} \{\|\mathbf{z}_t(\omega)\}\| \leq \sup_{t\in\mathbb{Z}} \left\{ \operatorname*{ess\,sup}_{\omega\in\Omega} \{\|\mathbf{z}_t(\omega)\|\} \right\}, \quad \text{almost surely,} \tag{6.21}$$

which is exactly what is required in (6.20) in order for (6.19) to hold. ■

## 6.13 Proof of Lemma 4.3

It is obvious that $L^\infty(\Omega, \ell^\infty(\mathbb{R}^n)) \subset L^\infty(\Omega, \ell^\infty(\mathbb{R}^n))$. Conversely, if $\mathbf{z} \in L^\infty(\Omega, \ell^\infty(\mathbb{R}^n))$ then, by definition, $\|\mathbf{z}\|_{L^\infty} < \infty$ or, equivalently, $\operatorname{ess\,sup}_{\omega\in\Omega} \{\sup_{t\in\mathbb{Z}} \{\|\mathbf{z}_t(\omega)\|\}\} < \infty$. By part **(i)** in Lemma 4.1, this implies that

$$\sup_{t\in\mathbb{Z}} \|\mathbf{z}_t(\omega)\| < \infty, \quad \text{almost surely.} \tag{6.22}$$

Since the elements in the spaces in (4.7) 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 $\mathbf{z}$ 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 (6.22) only in a set of zero measure, they are equal in $L^\infty\left(\Omega, (\mathbb{R}^n)^{\mathbb{Z}}\right)$ but, this time, $\mathbf{z}^* \in L^\infty(\Omega, \ell^\infty(\mathbb{R}^n))$, as required. Since $(\ell^\infty(\mathbb{R}^n), \|\cdot\|_\infty)$ is a Banach space, so is $(L^\infty(\Omega, \ell^\infty(\mathbb{R}^n)), \|\cdot\|_{L^\infty})$ by the references quoted after (4.2) and the statement follows. ■

## 6.14 Proof of Theorem 4.4

**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 2.2 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 $H : (K_M^{L^\infty}, \|\cdot\|_{L^\infty_w}) \longrightarrow L^\infty(\Omega, \mathbb{R})$ 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

$$\|\mathbf{z} - \mathbf{s}\|_w = \sup_{t \in \mathbb{Z}_-} \|(\mathbf{z}_t - \mathbf{s}_t) w_{-t}\| < \delta(\epsilon), \quad \text{then} \quad |H(\mathbf{z}) - H(\mathbf{s})| < \epsilon. \tag{6.23}$$

Moreover, since by Lemma 2.2 the space $(K_M, \|\cdot\|_w)$ is compact, the Uniform Continuity Theorem [Munk 14, Theorem 7.3] 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 $\|\mathbf{Z} - \mathbf{S}\|_{L^\infty_w} < \delta(\epsilon)$. Indeed, $\|\mathbf{Z} - \mathbf{S}\|_{L^\infty_w} < \delta(\epsilon)$ 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 4.1 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 (6.23), that $|H(\mathbf{Z}) - H(\mathbf{S})| < \epsilon$, almost surely, which by part **(ii)** in Lemma 4.1 implies that $\|H(\mathbf{Z}) - H(\mathbf{S})\|_{L^\infty} < \epsilon$, as required.

Conversely, if $H : (K_M^{L^\infty}, \|\cdot\|_{L^\infty_w}) \longrightarrow L^\infty(\Omega, \mathbb{R})$ 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 $H : (K_M^{L^\infty}, \|\cdot\|_{L^\infty_w}) \longrightarrow L^\infty(\Omega, \mathbb{R})$ be an arbitrary causal and time-invariant FMP filter and let $H_S \in \mathcal{T}$ 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 4.1 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 $H : (K_M^{L^\infty}, \|\cdot\|_{L^\infty_w}) \longrightarrow L^\infty(\Omega, \mathbb{R})$ 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 $H_S : (K_M^{L^\infty}, \|\cdot\|_{L^\infty_w}) \longrightarrow L^\infty(\Omega, \mathbb{R})$ be the element that, for a given $\epsilon > 0$, satisfies that $\|H - H_S\|_{L^\infty} = \sup_{\mathbf{Z} \in K_M^{L^\infty}} \|H(\mathbf{Z}) - H_S(\mathbf{Z})\|_{L^\infty} < \epsilon$. 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 Theorem 4.5

We first notice that the polynomial algebra $\mathcal{A}(\mathcal{R})$ is, by Theorem 3.1 and the first part of Theorem 4.4, 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 3.1 and it is hence universal in the deterministic setup. The result follows from the second part of Theorem 4.4. ■

**Acknowledgments:** We thank Philipp Harms for carefully looking at early versions of this work and for making suggestions that have significantly improved some of our results and Herbert Jaeger and Josef Teichmann for fruitful discussions. 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 lead 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).

## References

[Abra 88] R. Abraham, J. E. Marsden, and T. S. Ratiu. *Manifolds, Tensor Analysis, and Applications*. Vol. 75, Applied Mathematical Sciences. Springer-Verlag, 1988.

[Apos 74] T. Apostol. *Mathematical Analysis*. Addison Wesley, second Ed., 1974.

[Arno 57] V. I. Arnold. "On functions of three variables". *Proceedings of the USSR Academy of Sciences*, Vol. 114, pp. 679–681, 1957.

[Bai 12] Bai Zhang, D. J. Miller, and Yue Wang. "Nonlinear system modeling with random matrices: echo state networks revisited". *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 23, No. 1, pp. 175–182, jan 2012.

[Barr 93] A. Barron. "Universal approximation bounds for superpositions of a sigmoidal function". *IEEE Transactions on Information Theory*, Vol. 39, No. 3, pp. 930–945, may 1993.

[Boll 86] T. Bollerslev. "Generalized autoregressive conditional heteroskedasticity". *Journal of Econometrics*, Vol. 31, No. 3, pp. 307–327, 1986.

[Box 76] G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1976.

[Boyd 85] S. Boyd and L. Chua. "Fading memory and the problem of approximating nonlinear operators with Volterra series". *IEEE Transactions on Circuits and Systems*, Vol. 32, No. 11, pp. 1150–1161, nov 1985.

[Broc 06] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer-Verlag, 2006.

[Bueh 06] M. Buehner and P. Young. "A tighter bound for the echo state property". *IEEE Transactions on Neural Networks*, Vol. 17, No. 3, pp. 820–824, 2006.

[Come 06] F. Comets and T. Meyre. *Calcul Stochastique et Modèles de Diffusions*. Dunod, Paris, 2006.

[Coui 16] R. Couillet, G. Wainrib, H. Sevi, and H. T. Ali. "The asymptotic performance of linear echo state neural networks". *Journal of Machine Learning Research*, Vol. 17, No. 178, pp. 1–35, 2016.

[Croo 07] N. Crook. "Nonlinear transient computation". *Neurocomputing*, Vol. 70, pp. 1167–1176, 2007.

[Cybe 89] G. Cybenko. "Approximation by superpositions of a sigmoidal function". *Mathematics of Control, Signals, and Systems*, Vol. 2, No. 4, pp. 303–314, dec 1989.

[Damb 12] J. Dambre, D. Verstraeten, B. Schrauwen, and S. Massar. "Information processing capacity of dynamical systems". *Scientific reports*, Vol. 2, No. 514, 2012.

[Dieu 69] J. Dieudonne. *Foundations of Modern Analysis*. Academic Press, 1969.

[Doya 92] K. Doya. "Bifurcations in the learning of recurrent neural networks". In: *Proceedings of IEEE International Symposium on Circuits and Systems*, pp. 2777–2780, IEEE, 1992.

[Engl 82] R. F. Engle. "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation". *Econometrica*, Vol. 50, No. 4, pp. 987–1007, 1982.

[Flie 76] M. Fliess. "Un outil algebrique : les series formelles non commutatives". In: G. Marchesini and S. K. Mitter, Eds., *Mathematical Systems Theory*, pp. 122–148, Springer Verlag, 1976.

[Flie 80] M. Fliess and D. Normand-Cyrot. "Vers une approche algébrique des systèmes non linéaires en temps discret". In: A. Bensoussan and J. Lions, Eds., *Analysis and Optimization of Systems. Lecture Notes in Control and Information Sciences, vol. 28*, Springer Berlin Heidelberg, 1980.

[Fran 10] C. Francq and J.-M. Zakoian. *GARCH Models: Structure, Statistical Inference and Financial Applications*. Wiley, 2010.

[Galt 14] M. N. Galtier, C. Marini, G. Wainrib, and H. Jaeger. "Relative entropy minimizing noisy non-linear neural network to approximate stochastic processes". *Neural Networks*, Vol. 56, pp. 10–21, 2014.

[Gang 08] S. Ganguli, D. Huh, and H. Sompolinsky. "Memory traces in dynamical systems.". *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 105, No. 48, pp. 18970–5, dec 2008.

[Grig 15] L. Grigoryeva, J. Henriques, L. Larger, and J.-P. Ortega. "Optimal nonlinear information processing capacity in delay-based reservoir computers". *Scientific Reports*, Vol. 5, No. 12858, pp. 1–11, 2015.

[Grig 16a] L. Grigoryeva, J. Henriques, L. Larger, and J.-P. Ortega. "Nonlinear memory capacity of parallel time-delay reservoir computers in the processing of multidimensional signals". *Neural Computation*, Vol. 28, pp. 1411–1451, 2016.

[Grig 16b] L. Grigoryeva, J. Henriques, and J.-P. Ortega. "Reservoir computing: information processing of stationary signals". In: *Proceedings of the 19th IEEE International Conference on Computational Science and Engineering*, pp. 496–503, 2016.

[Grig 18] L. Grigoryeva and J.-P. Ortega. "Echo state networks are universal". *Preprint*, 2018.

[Grim 01] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, 2001.

[Herm 10] M. Hermans and B. Schrauwen. "Memory in linear recurrent neural networks in continuous time.". *Neural networks : the official journal of the International Neural Network Society*, Vol. 23, No. 3, pp. 341–55, apr 2010.

[Horn 13] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, second Ed., 2013.

[Horn 89] K. Hornik, M. Stinchcombe, and H. White. "Multilayer feedforward networks are universal approximators". *Neural Networks*, Vol. 2, No. 5, pp. 359–366, 1989.

[Jaeg 02] H. Jaeger. "Short term memory in echo state networks". *Fraunhofer Institute for Autonomous Intelligent Systems. Technical Report.*, Vol. 152, 2002.

[Jaeg 04] H. Jaeger and H. Haas. "Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication". *Science*, Vol. 304, No. 5667, pp. 78–80, 2004.

[Jaeg 10] H. Jaeger. "The 'echo state' approach to analysing and training recurrent neural networks with an erratum note". Tech. Rep., German National Research Center for Information Technology, 2010.

[Jone 92] 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*, Vol. 20, No. 1, pp. 608–613, 1992.

[Kolm 56] A. N. Kolmogorov. "On the representation of continuous functions of several variables as superpositions of functions of smaller number of variables". *Soviet Math. Dokl*, Vol. 108, pp. 179–182, 1956.

[Kurk 05] V. Kurkova and M. Sanguineti. "Learning with generalization capability by kernel methods of bounded complexity". *Journal of Complexity*, Vol. 21, No. 3, pp. 350–367, 2005.

[Ledo 91] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer-Verlag, 1991.

[Lord 14] G. J. Lord, C. E. Powell, and T. Shardlow. *An Introduction to Computational Stochastic PDES*. Cambridge University Press, 2014.

[Luko 09] M. Lukoševičius and H. Jaeger. "Reservoir computing approaches to recurrent neural network training". *Computer Science Review*, Vol. 3, No. 3, pp. 127–149, 2009.

[Maas 00] W. Maass and E. D. Sontag. "Neural Systems as Nonlinear Filters". *Neural Computation*, Vol. 12, No. 8, pp. 1743–1772, aug 2000.

[Maas 02] W. Maass, T. Natschläger, and H. Markram. "Real-time computing without stable states: a new framework for neural computation based on perturbations". *Neural Computation*, Vol. 14, pp. 2531–2560, 2002.

[Maas 04] W. Maass, T. Natschläger, and H. Markram. "Fading memory and kernel properties of generic cortical microcircuit models". *Journal of Physiology Paris*, Vol. 98, No. 4-6 SPEC. ISS., pp. 315–330, 2004.

[Maas 07] W. Maass, P. Joshi, and E. D. Sontag. "Computational aspects of feedback in neural circuits". *PLoS Computational Biology*, Vol. 3, No. 1, p. e165, 2007.

[Maas 11] W. Maass. "Liquid state machines: motivation, theory, and applications". In: S. S. Barry Cooper and A. Sorbi, Eds., *Computability In Context: Computation and Logic in the Real World*, Chap. 8, pp. 275–296, 2011.

[Mack 77] M. C. Mackey and L. Glass. "Oscillation and chaos in physiological control systems". *Science*, Vol. 197, pp. 287–289, 1977.

[Manj 13] G. Manjunath and H. Jaeger. "Echo state property linked to an input: exploring a fundamental characteristic of recurrent neural networks". *Neural Computation*, Vol. 25, No. 3, pp. 671–696, 2013.

[Matt 92] M. B. Matthews. *On the UniformApproximation of Nonlinear Discrete-Time Fading-Memory Systems Using Neural Network Models*. PhD thesis, ETH Zürich, 1992.

[Matt 93] M. B. Matthews. "Approximating nonlinear fading-memory operators using neural network models". *Circuits, Systems, and Signal Processing*, Vol. 12, No. 2, pp. 279–307, jun 1993.

[Munk 14] J. Munkres. *Topology*. Pearson, second Ed., 2014.

[Perr 96] P. C. Perryman. *Approximation Theory for Deterministic and Stochastic Nonlinear Systems*. PhD thesis, University of California, Irvine, 1996.

[Perr 97] 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*, pp. 146–150, IEEE Comput. Soc. Press, 1997.

[Pisi 81] G. Pisier. "Remarques sur un résultat non publié de B. Maurey". *Séminaire d'analyse fonctionnelle École Polytechnique*, pp. 1–12, 1981.

[Rusc 98] L. Rüschendorf and W. Thomsen. "Closedness of sum spaces and the generalized schrödinger Problem". *Theory of Probability & Its Applications*, Vol. 42, No. 3, pp. 483–494, jan 1998.

[Sont 79a] E. Sontag. "Realization theory of discrete-time nonlinear systems: Part I-The bounded case". *IEEE Transactions on Circuits and Systems*, Vol. 26, No. 5, pp. 342–356, may 1979.

[Sont 79b] E. D. Sontag. "Polynomial Response Maps". In: *Lecture Notes Control in Control and Information Sciences. Vol. 13*, Springer Verlag, 1979.

[Spre 65] D. A. Sprecher. "A representation theorem for continuous functions of several variables". *Proceedings of the American Mathematical Society*, Vol. 16, No. 2, p. 200, apr 1965.

[Spre 96] D. A. Sprecher. "A numerical implementation of Kolmogorov's superpositions". *Neural Networks*, Vol. 9, No. 5, pp. 765–772, 1996.

[Spre 97] D. A. Sprecher. "A numerical implementation of Kolmogorov's superpositions II". *Neural Networks*, Vol. 10, No. 3, pp. 447–457, 1997.

[Stub 97a] 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*, pp. 141–145, IEEE Comput. Soc. Press, 1997.

[Stub 97b] A. Stubberud and P. Perryman. "State of system approximation for stochastic systems". In: *Proceedings of 13th International Conference on Digital Signal Processing*, pp. 711–714, IEEE, 1997.

[Suss 76] H. J. Sussmann. "Semigroup representations, bilinear approximations of input-output maps, and generalized inputs". In: G. Marchesini and S. K. MItter, Eds., *Mathematical Systems Theory*, pp. 172–191, Springer Verlag, 1976.

[Take 81] F. Takens. "Detecting strange attractors in turbulence". pp. 366–381, Springer Berlin Heidelberg, 1981.

[Vers 07] D. Verstraeten, B. Schrauwen, M. D'Haene, and D. Stroobandt. "An experimental unification of reservoir computing methods". *Neural Networks*, Vol. 20, pp. 391–403, 2007.

[Wain 16] G. Wainrib and M. N. Galtier. "A local echo state property through the largest Lyapunov exponent". *Neural Networks*, Vol. 76, pp. 39–45, apr 2016.

[Whit 04] O. White, D. Lee, and H. Sompolinsky. "Short-Term Memory in Orthogonal Neural Networks". *Physical Review Letters*, Vol. 92, No. 14, p. 148102, apr 2004.

[Yild 12] I. B. Yildiz, H. Jaeger, and S. J. Kiebel. "Re-visiting the echo state property.". *Neural networks : the official journal of the International Neural Network Society*, Vol. 35, pp. 1–9, nov 2012.

[Zang 04] G. Zang and P. A. Iglesias. "Fading memory and stability". *Journal of the Franklin Institute*, Vol. 340, No. 6-7, pp. 489–502, 2004.