Reservoir Computing Universality With Stochastic Inputs

Lukas Gonon and Juan-Pablo Ortega

***Abstract*—The universal approximation properties with respect to L^p-type criteria of three important families of reservoir computers with stochastic discrete-time semi-infinite inputs is shown. First, it is proved that linear reservoir systems with either polynomial or neural network readout maps are universal. More importantly, it is proved that the same property holds for two families with linear readouts, namely, trigonometric state-affine systems and echo state networks, which are the most widely used reservoir systems in applications. The linearity in the readouts is a key feature in supervised machine learning applications. It guarantees that these systems can be used in high-dimensional situations and in the presence of large datasets. The L^p criteria used in this paper allow the formulation of universality results that do not necessarily impose almost sure uniform boundedness in the inputs or the fading memory property in the filter that needs to be approximated.**

***Index Terms*—Reservoir computing, echo state network, ESN, machine learning, uniform system approximation, stochastic input, universality.**

I. Introduction

A UNIVERSALITY statement in relation to a machine learning paradigm refers to its versatility at the time of reproducing a rich number of patterns obtained by modifying only a limited number of hyperparameters. In the language of learning theory, universality amounts to the possibility of making approximation errors as small as one wants [1]–[3]. Well-known universality results are, for example, the uniform approximation properties of feedforward neural networks established in [4], [5] for deterministic inputs and, later on, extended in [6] to accommodate random inputs.

This paper is a generalization of the universality statements in [6] to a discrete-time dynamical context. More specifically, we are interested in the learning not of functions but of filters that transform semi-infinite random input sequences parameterized by time into outputs that depend on those inputs in a causal and time-invariant manner. The approximants used are small subfamilies of reservoir computers (RC) [7], [8] or reservoir systems. Reservoir computers are filters generated by nonlinear state-space transformations and constitute special types of recurrent neural networks. They are 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 a state-space transformation given by:

$$\begin{cases} \mathbf{x}_t = F(\mathbf{x}_{t-1}, \mathbf{z}_t), & (1) \\ y_t = h(\mathbf{x}_t), & (2) \end{cases}$$

where $t \in \mathbb{Z}$ and the dimension $N \in \mathbb{N}$ of the state vectors $\mathbf{x}_t \in \mathbb{R}^N$ is referred to as the number of virtual neurons of the system. In supervised machine learning applications the reservoir map is very often randomly generated and the memoryless readout is trained so that the output matches a given teaching signal.

Families of systems of this type have already been proved to be universal in different contexts. In the continuous-time setup, it was shown in [9] that linear reservoir systems with polynomial readouts or bilinear reservoirs with linear readouts are able to uniformly approximate any fading memory filter with uniformly bounded and equicontinuous inputs. The fading memory property is a continuity feature exhibited by many filters encountered in applications.

In the discrete-time setup, several universality statements were already part of classical systems theory statements for inputs defined on a finite number of time points [10]–[12]. In the more general context of semi-infinite inputs, various universality results have been formulated for systems with approximate finite memory [13]–[18]. These universality results have been recently extended to the causal and fading memory category in [19], [20]. In those works it has been established the universality of two important families of reservoir systems with linear readouts, namely, the so called state affine systems (SAS) and the echo state networks (ESN). Moreover, the universality of the SAS family was established in [19] both for uniformly bounded deterministic inputs, as well as for almost surely uniformly bounded stochastic ones. This last statement was shown to be a corollary of a general transfer theorem that proves that very important features of causal and time-invariant filters like the fading memory property or universality are naturally inherited by reservoir systems with almost surely uniformly bounded stochastic inputs from their counterparts with deterministic inputs.

Unfortunately, almost surely bounded random inputs are not always appropriate for many applications. For example, most parametric time series models use as driving innovations random variables whose distributions are not compactly supported (Gaussian, for example) in order to ensure adequate levels of performance. The main goal of this work is *formulating universality results in the stochastic context that do not impose almost sure uniform boundedness in the inputs*.

L. Gonon and J.-P. Ortega are with the Department of Mathematics and Statistics, Universität Sankt Gallen, Sankt Gallen, Switzerland. L. Gonon is also affiliated with the Department of Mathematics, ETH Zürich, Switzerland. J.-P. Ortega is also affiliated with the Centre National de la Recherche Scientifique (CNRS), France.

The way in which the universality results contained in this paper are articulated differs somewhat from the above quoted references and is more in the vein of [6]. More specifically, in the stochastic universality statements in [19], for example, universal families are presented that uniformly approximate any given filter for any input in a given class of stochastic processes. In contrast with this strategy and like in [6], we fix here first a discrete-time stochastic process that models the data generating process (DGP) behind the system inputs that are being considered. Subsequently, families of reservoir filters are spelled out whose images of the DGP are dense in the L^p sense. Equivalently, the image of the DGP by any measurable causal and time invariant filter can be approximated by the image of one of the members of the universal family with respect to an L^p norm defined using the law of the prefixed DGP.

It is important to point out that this approach allows us to *formulate universality results for filters that do not necessarily have the fading memory property since only measurability is imposed as a hypothesis*.

The paper contains three main universality statements. The first one shows that linear reservoir systems with either polynomial or neural network readout maps are universal in the L^p sense. More importantly, two other families with linear readouts are shown to also have this property, namely, trigonometric state-affine systems and echo state networks, which are the most widely used reservoir systems in applications. The linearity of the readout is a key feature of these systems since in supervised machine learning applications it reduces the training task to the solution of a linear regression problem, which can be implemented efficiently also in high-dimensional situations and in the presence of large datasets.

II. Preliminaries

In this section we introduce some notation and collect general facts about filters, reservoir systems, and stochastic input signals.

A. Notation

We write $\mathbb{N} = \{0, 1, \ldots\}$ and $\mathbb{Z}_- = \{\ldots, -1, 0\}$. The elements of the Euclidean spaces $\mathbb{R}^n$ will be written as column vectors and will be denoted in bold. Given a vector $\mathbf{v} \in \mathbb{R}^n$, we denote its entries by v_i or by $v^{(i)}$, with $i \in \{1, \ldots, n\}$. $(\mathbb{R}^n)^{\mathbb{Z}}$ and $(\mathbb{R}^n)^{\mathbb{Z}_-}$ denote the sets of infinite $\mathbb{R}^n$-valued sequences of the type $(\ldots, \mathbf{z}_{-1}, \mathbf{z}_0, \mathbf{z}_1, \ldots)$ and $(\ldots, \mathbf{z}_{-1}, \mathbf{z}_0)$ with $\mathbf{z}_i \in \mathbb{R}^n$ for $i \in \mathbb{Z}$ and $i \in \mathbb{Z}_-$, respectively. The elements in these sequence spaces will also be written in bold, for example, $\mathbf{z} := (\ldots, \mathbf{z}_{-1}, \mathbf{z}_0) \in (\mathbb{R}^n)^{\mathbb{Z}_-}$. 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. Random variables and stochastic processes will be denoted using upper case characters that will be bold when they are vector valued.

B. Filters and functionals

A filter is a map $U\colon (\mathbb{R}^n)^{\mathbb{Z}} \to \mathbb{R}^{\mathbb{Z}}$. It is called causal, if for any $\mathbf{z}, \mathbf{w} \in (\mathbb{R}^n)^{\mathbb{Z}}$ which satisfy $\mathbf{z}_\tau = \mathbf{w}_\tau$ for all $\tau \leq t$ for a given $t \in \mathbb{Z}$, one has that $U(\mathbf{z})_t = U(\mathbf{w})_t$. Denote by $T_{-\tau}\colon (\mathbb{R}^n)^{\mathbb{Z}} \to (\mathbb{R}^n)^{\mathbb{Z}}$ the time delay operator defined by $T_{-\tau}(\mathbf{z})_t := \mathbf{z}_{t+\tau}$, for any $\tau \in \mathbb{Z}$. A filter U is called time-invariant, if $T_{-\tau} \circ U = U \circ T_{-\tau}$ for all $\tau \in \mathbb{Z}$.

Causal and time-invariant filters can be equivalently described using their naturally associated functionals. We refer to a map $H\colon (\mathbb{R}^n)^{\mathbb{Z}_-} \to \mathbb{R}$ as a functional. Given a causal and time-invariant filter U, one defines the functional H_U associated to it by setting $H_U(\mathbf{z}) := U(\mathbf{z}^e)_0$. Here $\mathbf{z}^e$ is an arbitrary extension of $\mathbf{z} \in (\mathbb{R}^n)^{\mathbb{Z}_-}$ to $(\mathbb{R}^n)^{\mathbb{Z}}$. H_U does not depend on the choice of this extension since U is causal. Conversely, given a functional H one may define a causal and time-invariant filter $U_H\colon (\mathbb{R}^n)^{\mathbb{Z}} \to \mathbb{R}^{\mathbb{Z}}$ by setting $U_H(\mathbf{z})_t := H(\pi_{\mathbb{Z}_-} \circ T_{-t}(\mathbf{z}))$, where $\pi_{\mathbb{Z}_-}\colon (\mathbb{R}^n)^{\mathbb{Z}} \to (\mathbb{R}^n)^{\mathbb{Z}_-}$ is the natural projection. One may verify that any causal and time-invariant filter can be recovered from its associated functional and conversely. Equivalently, $U = U_{H_U}$ and $H = H_{U_H}$. We refer to [9] for further details.

If U is causal and time-invariant, then for any $\mathbf{z} \in (\mathbb{R}^n)^{\mathbb{Z}}$ the sequence $U(\mathbf{z})$ restricted to $\mathbb{Z}_-$ only depends on $(\mathbf{z}_t)_{t\in\mathbb{Z}_-}$. Thus we may also consider U as a map $U\colon (\mathbb{R}^n)^{\mathbb{Z}_-} \to \mathbb{R}^{\mathbb{Z}_-}$, but when we do so this will always be clear from the context.

C. Reservoir computing systems

A specific class of filters can be obtained using the reservoir computing systems or reservoir computers (RC) introduced in (1)-(2) when they satisfy the following property: a reservoir system satisfies the echo state property (ESP) if for any $\mathbf{z} \in (\mathbb{R}^n)^{\mathbb{Z}}$ there exists a unique $\mathbf{x} \in (\mathbb{R}^N)^{\mathbb{Z}}$ such that (1) holds. In this case the RC system gives rise to a filter U_h^F associating to any $\mathbf{z} \in (\mathbb{R}^n)^{\mathbb{Z}}$ the unique output in (2), that is, $U_h^F(\mathbf{z})_t := y_t$. Furthermore, it can be shown (see [20, Proposition 2.1]) that U_h^F is necessarily causal and time-invariant and hence we may associate to U_h^F a reservoir functional $H_h^F\colon (\mathbb{R}^n)^{\mathbb{Z}_-} \to \mathbb{R}$ defined as $H_h^F(\mathbf{z}) := U_h^F(\mathbf{z})_0$.

As seen above, the causal and time-invariant filter U_h^F is uniquely determined by the reservoir functional H_h^F. Since the latter is determined by the restriction of the RC system to $\mathbb{Z}_-$, we will sometimes consider the system (1)-(2) only for $t \in \mathbb{Z}_-$.

D. Deterministic filters with stochastic inputs

We are interested in feeding the filters and the systems that we just introduced with stochastic processes as inputs. More explicitly, given a causal and time-invariant filter U that satisfies certain measurability hypotheses, any stochastic process $\mathbf{Z} = (\mathbf{Z}_t)_{t\in\mathbb{Z}_-}$ is mapped to a new stochastic process $(U(\mathbf{Z})_t)_{t\in\mathbb{Z}_-}$. The main contributions in this article address the question of approximating $U(\mathbf{Z})$ by reservoir filters in an L^p-sense. We now introduce the precise framework to achieve this goal.

1) Probabilistic framework: Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which all random variables are defined. The input signal is modeled as a discrete-time stochastic process $\mathbf{Z} = (\mathbf{Z}_t)_{t\in\mathbb{Z}_-}$ with values in $\mathbb{R}^n$. When dealing with stochastic processes we will make no distinctions between the assignment $\mathbf{Z} : \mathbb{Z}_- \times \Omega \to \mathbb{R}^n$ and the corresponding map into

path space $\mathbf{Z}:\Omega\to(\mathbb{R}^n)^{\mathbb{Z}_-}$. We recall that $\mathbf{Z}$ is a stochastic process when the corresponding map $\mathbf{Z}:\Omega\to(\mathbb{R}^n)^{\mathbb{Z}_-}$ is measurable. Here $(\mathbb{R}^n)^{\mathbb{Z}_-}$ is equipped with the product σ-algebra $\otimes_{t\in\mathbb{Z}_-}\mathcal{B}(\mathbb{R}^n)$ (which coincides with the Borel σ-algebra of $(\mathbb{R}^n)^{\mathbb{Z}_-}$ equipped with the product topology by [21, Lemma 1.2]), where $\mathcal{B}(\mathbb{R}^n)$ is the Borel σ-algebra on $\mathbb{R}^n$.

We denote by $\mathcal{F}_t:=\sigma(\mathbf{Z}_0,\dots,\mathbf{Z}_t)$, $t\in\mathbb{Z}_-$, the σ-algebra generated by $\{\mathbf{Z}_0,\dots,\mathbf{Z}_t\}$ and write $\mathcal{F}_{-\infty}:=\sigma(\mathbf{Z}_t\colon t\in\mathbb{Z}_-)$. For $p\in[1,\infty]$ we denote by $L^p(\Omega,\mathcal{F},\mathbb{P})$ the Banach space formed by the real-valued random variables in $(\Omega,\mathcal{F},\mathbb{P})$ that have a finite usual L^p norm $\|\cdot\|_p$.

We say that the process $\mathbf{Z}$ is stationary when for any $\{t_1,\dots,t_k\}\subset\mathbb{Z}_-$, $h\in\mathbb{Z}_-$, and $A_{t_1},\dots,A_{t_k}\in\mathcal{B}(\mathbb{R}^n)$, we have that

$$\begin{aligned}&\mathbb{P}\left(\mathbf{Z}_{t_1}\in A_{t_1},\dots,\mathbf{Z}_{t_k}\in A_{t_k}\right)\\&\qquad=\mathbb{P}\left(\mathbf{Z}_{t_1+h}\in A_{t_1},\dots,\mathbf{Z}_{t_k+h}\in A_{t_k}\right).\end{aligned}$$

2) Measurable functionals and filters: We say that a functional H is measurable when the map between measurable spaces $H\colon\left((\mathbb{R}^n)^{\mathbb{Z}_-},\otimes_{t\in\mathbb{Z}_-}\mathcal{B}(\mathbb{R}^n)\right)\to(\mathbb{R},\mathcal{B}(\mathbb{R}))$ is measurable. When H is measurable then so is $H(\mathbf{Z}):(\Omega,\mathcal{F})\to(\mathbb{R},\mathcal{B}(\mathbb{R}))$ since $H(\mathbf{Z})=H\circ\mathbf{Z}$ is the composition of measurable maps and hence $H(\mathbf{Z})$ is a random variable on $(\Omega,\mathcal{F},\mathbb{P})$.

Analogously, we will say that a causal, time-invariant filter U is measurable when the map between measurable spaces $U\colon\left((\mathbb{R}^n)^{\mathbb{Z}},\otimes_{t\in\mathbb{Z}}\mathcal{B}(\mathbb{R}^n)\right)\to\left(\mathbb{R}^{\mathbb{Z}},\otimes_{t\in\mathbb{Z}}\mathcal{B}(\mathbb{R})\right)$ is measurable. In that case, also the restriction of U to $\mathbb{Z}_-$ (see above) is measurable and so $U(\mathbf{Z})$ is a real-valued stochastic process.

As discussed above, causal, time-invariant filters and functionals are in a one-to-one correspondence. This relation is compatible with the measurability condition, that is, a causal and time-invariant filter is measurable if and only if the associated functional is measurable. In order to prove this statement we show first that the operator $\pi_{\mathbb{Z}_-}\circ T_{-t}:\left((\mathbb{R}^n)^{\mathbb{Z}},\otimes_{t\in\mathbb{Z}}\mathcal{B}(\mathbb{R}^n)\right)\longrightarrow\left((\mathbb{R}^n)^{\mathbb{Z}_-},\otimes_{t\in\mathbb{Z}_-}\mathcal{B}(\mathbb{R}^n)\right)$ is a measurable map, for any $t\in\mathbb{Z}_-$. Indeed, notice first that the projections $p_i:\left((\mathbb{R}^n)^{\mathbb{Z}},\otimes_{t\in\mathbb{Z}}\mathcal{B}(\mathbb{R}^n)\right)\longrightarrow(\mathbb{R}^n,\mathcal{B}(\mathbb{R}^n))$, $i\in\mathbb{Z}_-$, given by $p_i(\mathbf{z})=\mathbf{z}_i$ are measurable. Since $\pi_{\mathbb{Z}_-}\circ T_{-t}$ can be written as the Cartesian product of measurable maps $\pi_{\mathbb{Z}_-}\circ T_{-t}=\prod_{i=-\infty}^{t}p_i=(\dots,p_{t-2},p_{t-1},p_t)$, it is hence measurable [21, Lemma 1.8].

Now, if H is a measurable functional, this implies that the associated filter

$$U_H=\prod_{t=-\infty}^{0}H\circ\pi_{\mathbb{Z}_-}\circ T_{-t}\tag{3}$$

is also measurable since it is a composition of measurable functions. Conversely, if U is causal, time-invariant, and measurable, then so is the associated functional $H_U=p_0\circ U$.

3) L^p-norm for functionals: Fix $p\in[1,\infty)$ and let H be a measurable functional such that $H(\mathbf{Z})\in L^p(\Omega,\mathcal{F},\mathbb{P})$. The functionals which satisfy that

$$\|H(\mathbf{Z})\|_p:=\mathbb{E}[|H(\mathbf{Z})|^p]^{1/p}<\infty\tag{4}$$

will be referred to as p-integrable with respect to the input process $\mathbf{Z}$.

Let us now consider the expression (4) from an alternative point of view. Denote by $\mu_{\mathbf{Z}}:=\mathbb{P}\circ\mathbf{Z}^{-1}$ the law of $\mathbf{Z}$ when viewed as a $(\mathbb{R}^n)^{\mathbb{Z}_-}$-valued random variable as above. Thus $\mu_{\mathbf{Z}}$ is a probability measure on $(\mathbb{R}^n)^{\mathbb{Z}_-}$ such that for any measurable set $A\subset(\mathbb{R}^n)^{\mathbb{Z}_-}$ one has $\mu_{\mathbf{Z}}(A)=\mathbb{P}(\mathbf{Z}\in A)$. The requirement $H(\mathbf{Z})\in L^p(\Omega,\mathcal{F},\mathbb{P})$ then translates to $H\in L^p((\mathbb{R}^n)^{\mathbb{Z}_-},\mu_{\mathbf{Z}})$ and (4) is equal [21, Lemma 1.22] to

$$\|H\|_p^{\mu_{\mathbf{Z}}}:=\left[\int_{(\mathbb{R}^n)^{\mathbb{Z}_-}}|H(\mathbf{z})|^p\mu_{\mathbf{Z}}(\mathrm{d}\mathbf{z})\right]^{1/p}=\|H(\mathbf{Z})\|_p.$$

Thus, the results formulated later on in the paper for functionals with random inputs can also be seen as statements for functionals with deterministic inputs in $(\mathbb{R}^n)^{\mathbb{Z}_-}$, where the closeness between them is measured using the norm in $L^p((\mathbb{R}^n)^{\mathbb{Z}_-},\mu_{\mathbf{Z}})$. Following the terminology used by [6] we will refer to $\mu_{\mathbf{Z}}$ as the input environment measure.

We emphasize that these two points of view are equivalent. Given any probability measure $\mu_{\mathbf{Z}}$ on $(\mathbb{R}^n)^{\mathbb{Z}_-}$ one may set $\Omega=(\mathbb{R}^n)^{\mathbb{Z}_-}$, $\mathcal{F}=\otimes_{t\in\mathbb{Z}_-}\mathcal{B}(\mathbb{R}^n)$, $\mathbb{P}=\mu_{\mathbf{Z}}$ and define $Z_t(\mathbf{z}):=\mathbf{z}_t$ for all $\mathbf{z}\in\Omega$. We will switch between these two viewpoints throughout the paper without much warning to the reader.

4) L^p-norm for filters: Fix $p\in[1,\infty)$. A causal, time-invariant, measurable filter U is said to be p-integrable, if

$$\|U(\mathbf{Z})\|_p:=\sup_{t\in\mathbb{Z}_-}\left\{\mathbb{E}\left[|U(\mathbf{Z})_t|^p\right]^{1/p}\right\}<\infty.\tag{5}$$

It is easy to see that if U is p-integrable, then so is the corresponding functional H_U due to the following inequality

$$\begin{aligned}\|H_U(\mathbf{Z})\|_p&=\mathbb{E}[|H_U(\mathbf{Z})|^p]^{1/p}=\mathbb{E}[|U(\mathbf{Z})_0|^p]^{1/p}\\&\le\sup_{t\in\mathbb{Z}_-}\left\{\mathbb{E}\left[|U(\mathbf{Z})_t|^p\right]^{1/p}\right\}=\|U(\mathbf{Z})\|_p<\infty.\end{aligned}$$

The converse implication holds true when the input process is stationary. In order to show this fact, notice first that if μ_t is the law of $\pi_{\mathbb{Z}_-}\circ T_{-t}(\mathbf{Z})$, $t\in\mathbb{Z}_-$, and $\mathbf{Z}$ is by hypothesis stationary then, for any $\{t_1,\dots,t_k\}\subset\mathbb{Z}_-$ and $A_{t_1},\dots,A_{t_k}\in\mathcal{B}(\mathbb{R}^n)$, we have that

$$\begin{aligned}&\mathbb{P}\left((\pi_{\mathbb{Z}_-}\circ T_{-t}(\mathbf{Z}))_{t_1}\in A_{t_1},\dots,(\pi_{\mathbb{Z}_-}\circ T_{-t}(\mathbf{Z}))_{t_k}\in A_{t_k}\right)\\&\qquad=\mathbb{P}\left(\mathbf{Z}_{t_1+t}\in A_{t_1},\dots,\mathbf{Z}_{t_k+t}\in A_{t_k}\right)\\&\qquad\qquad=\mathbb{P}\left(\mathbf{Z}_{t_1}\in A_{t_1},\dots,\mathbf{Z}_{t_k}\in A_{t_k}\right),\end{aligned}$$

which proves that

$$\mu_{\mathbf{Z}}=\mu_t,\quad\text{for all}\quad t\in\mathbb{Z}_-.\tag{6}$$

This identity, together with (3), implies that for any p-integrable functional H:

$$\begin{aligned}\|U_H(\mathbf{Z})\|_p&=\sup_{t\in\mathbb{Z}_-}\left\{\mathbb{E}\left[|U_H(\mathbf{Z})_t|^p\right]^{1/p}\right\}\\&=\sup_{t\in\mathbb{Z}_-}\left\{\mathbb{E}\left[|H(\pi_{\mathbb{Z}_-}\circ T_{-t}(\mathbf{Z}))|^p\right]^{1/p}\right\}\\&=\sup_{t\in\mathbb{Z}_-}\left\{\left[\int_{(\mathbb{R}^n)^{\mathbb{Z}_-}}|H(\mathbf{z})|^p\mu_t(\mathrm{d}\mathbf{z})\right]^{1/p}\right\}\end{aligned}$$

$$= \sup_{t\in\mathbb{Z}_-}\left\{\left[\int_{(\mathbb{R}^n)^{\mathbb{Z}_-}}|H(\mathbf{z})|^p\mu_{\mathbf{Z}}(\mathrm{d}\mathbf{z})\right]^{1/p}\right\} = \|H(\mathbf{Z})\|_p < \infty, \tag{7}$$

which proves the p-integrability of the associated filter U_H.

III. L^p-UNIVERSALITY RESULTS

Fix $p \in [1,\infty)$, $\mathbf{Z}$ an input process, and a functional H such that $H(\mathbf{Z}) \in L^p(\Omega,\mathcal{F},\mathbb{P})$. The goal of this section is finding simple families of reservoir systems that are able to approximate $H(\mathbf{Z})$ as accurately as needed in the L^p-sense. The first part contains a result that shows that linear reservoir maps with polynomial readouts are able to carry this out. The situation is hence identical to the case for deterministic inputs or for almost surely uniformly bounded stochastic ones [19]. The second part contains a family that is able to achieve universality using only linear readouts, which is of major importance for applications since in that case the training effort reduces to solving a linear regression. Finally, we prove the universality of echo state networks which is the most widely used family of reservoir systems with linear readouts.

A. Linear reservoirs with nonlinear readouts

Consider a reservoir system with linear reservoir map and a polynomial readout. More precisely, given $A \in \mathbb{M}_N$, $\mathbf{c} \in \mathbb{M}_{N,n}$, and $h \in \mathrm{Pol}_N$ a real-valued polynomial in N variables, consider the system

$$\begin{cases}\mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{c}\mathbf{z}_t, & t\in\mathbb{Z}_-,\\ y_t = h(\mathbf{x}_t), & t\in\mathbb{Z}_-,\end{cases} \tag{8}$$

for any $\mathbf{z} \in (\mathbb{R}^n)^{\mathbb{Z}_-}$. If the matrix A is chosen so that $\sigma_{\max}(A) < 1$, then this system has the echo state property and the corresponding reservoir filter $U_h^{A,\mathbf{c}}$ is causal and time-invariant [19]. We denote by $H_h^{A,\mathbf{c}}$ the associated functional. We are interested in the approximation capabilities that can be achieved by using processes of the type $H_h^{A,\mathbf{c}}(\mathbf{Z})$, where $\mathbf{Z}$ is a fixed input process and $H_h^{A,\mathbf{c}}(\mathbf{Z}) = Y_0$, with Y_0 obviously determined by the stochastic reservoir system

$$\begin{cases}\mathbf{X}_t = A\mathbf{X}_{t-1} + c\mathbf{Z}_t, & t\in\mathbb{Z}_-,\\ Y_t = h(\mathbf{X}_t), & t\in\mathbb{Z}_-.\end{cases} \tag{9}$$

Proposition III.1. *Fix $p \in [1,\infty)$, let $\mathbf{Z}$ be a fixed $\mathbb{R}^n$-valued input process, and let H be a functional such that $H(\mathbf{Z}) \in L^p(\Omega,\mathcal{F},\mathbb{P})$. Suppose that for any $K\in\mathbb{N}$ there exists $\alpha > 0$ such that*

$$\mathbb{E}\left[\exp\left(\alpha\sum_{k=0}^{K}\sum_{i=1}^{n}|Z_{-k}^{(i)}|\right)\right] < \infty. \tag{10}$$

Then, for any $\varepsilon > 0$ there exists $N\in\mathbb{N}$, $A\in\mathbb{M}_N$, $\mathbf{c}\in\mathbb{M}_{N,n}$, and $h\in\mathrm{Pol}_N$ such that (8) has the echo state property, the corresponding filter is causal and time-invariant, the associated functional satisfies $H_h^{A,c}(\mathbf{Z}) \in L^p(\Omega,\mathcal{F},\mathbb{P})$, and

$$\|H(\mathbf{Z}) - H_h^{A,c}(\mathbf{Z})\|_p < \varepsilon. \tag{11}$$

If the input process $\mathbf{Z}$ is stationary then

$$\|U_H(\mathbf{Z}) - U_h^{A,c}(\mathbf{Z})\|_p < \varepsilon. \tag{12}$$

Proof. The proof consists of two steps: In the first one we use assumption (10) and classical results in the literature to establish that

$$\mathrm{Pol}_{n(K+1)} \text{ is dense in } L^p(\mathbb{R}^{n(K+1)},\mu_K), \text{ for all } K\in\mathbb{N}, \tag{13}$$

where μ_K is the law of $(Z_0^{(1)}, Z_0^{(2)},\dots,Z_{-K}^{(n-1)},Z_{-K}^{(n)})$ on $\mathbb{R}^{n(K+1)}$ under $\mathbb{P}$. In the second step we then use (13) to construct a linear RC system of the type in (8) that yields the approximation statement (11).

Step 1: Denote by μ_K the law of $(Z_0^{(1)}, Z_0^{(2)},\dots,Z_{-K}^{(n-1)},Z_{-K}^{(n)})$ on $\mathbb{R}^N$ under $\mathbb{P}$, where $N := n(K+1)$. By (10) there exists $\alpha > 0$ such that $\int_{\mathbb{R}^N}\exp(\alpha\|\mathbf{z}\|_1)\mu_K(\mathrm{d}\mathbf{z}) < \infty$, where here and in the rest of this proof $\|\cdot\|_1$ denotes the Euclidean 1-norm. Denoting by μ_K^j the j-th marginal distribution of μ_K, this implies for $j = 1,\dots,N$ that

$$\int_{\mathbb{R}}\exp(\alpha|z^{(j)}|)\mu_K^j(\mathrm{d}z^{(j)}) \le \int_{\mathbb{R}^N}\exp(\alpha\|\mathbf{z}\|_1)\mu_K(\mathrm{d}\mathbf{z}) < \infty.$$

Consequently, by [22, Theorem 6], Pol_1 is dense in $L^p(\mathbb{R},\mu_K^j)$ for any $p\in[1,\infty)$, $j=1,\dots,N$. By [23, Proposition page 364] this implies that Pol_N is dense in $L^p(\mathbb{R}^N,\mu_K)$, where we note that μ_K indeed satisfies the moment assumption in [23, Page 361]: since $x^{2m}\le\exp(\alpha x)$ for any $x\ge 0$, $m\in\mathbb{N}$, one has

$$\begin{aligned}\int_{\mathbb{R}^N}\|\mathbf{z}\|_2^{2m}\mu_K(\mathrm{d}\mathbf{z}) &\le \int_{\mathbb{R}^N}\exp(\alpha\|\mathbf{z}\|_2)\mu_K(\mathrm{d}\mathbf{z})\\ &\le \int_{\mathbb{R}^N}\exp(\alpha\|\mathbf{z}\|_1)\mu_K(\mathrm{d}\mathbf{z}) < \infty.\end{aligned}$$

Step 2: Let $\varepsilon > 0$. By Lemma A.1 in the appendix there exists $K\in\mathbb{N}$ such that

$$\|H(\mathbf{Z}) - \mathbb{E}[H(\mathbf{Z})|\mathcal{F}_{-K}]\|_p < \frac{\varepsilon}{2} \tag{14}$$

where $\mathcal{F}_{-K} := \sigma(\mathbf{Z}_0,\dots,\mathbf{Z}_{-K})$. In the following paragraphs we will establish the approximation statement (11) for $\mathbb{E}[H(Z)|\mathcal{F}_K]$ instead of $H(Z)$. Combining this with (14) will then yield (11).

Let $N := n(K+1)$. By definition, $\mathbb{E}[H(\mathbf{Z})|\mathcal{F}_{-K}]$ is $\mathcal{F}_{-K}$-measurable and hence there exists [21, Lemma 1.13] a measurable function $g_K\colon\mathbb{R}^N\to\mathbb{R}$ such that $\mathbb{E}[H(\mathbf{Z})|\mathcal{F}_{-K}] = g_K(\mathbf{Z}_0,\dots,\mathbf{Z}_{-K})$. Furthermore,

$$\begin{aligned}&\int_{\mathbb{R}^N}|g_K(\mathbf{z})|^p\mu_K(\mathrm{d}\mathbf{z})\\ &\qquad= \mathbb{E}[|\mathbb{E}[H(\mathbf{Z})|\mathcal{F}_{-K}]|^p] \le \mathbb{E}[|H(\mathbf{Z})|^p] < \infty,\end{aligned}$$

by standard properties of conditional expectations (see, for instance, [24, Theorem 5.1.4]) and the assumption that $H(\mathbf{Z})\in L^p(\Omega,\mathcal{F},\mathbb{P})$. Thus, $g_K\in L^p(\mathbb{R}^N,\mu_K)$ and using the statement (13) established in Step 1, there exists $h\in\mathrm{Pol}_N$ such that

$$\begin{aligned}&\|\mathbb{E}[H(\mathbf{Z})|\mathcal{F}_{-K}] - h(\mathbf{Z}_0^\top,\dots,\mathbf{Z}_{-K}^\top)\|_p\\ &\qquad= \|g_K - h\|_{L^p(\mathbb{R}^N,\mu_K)} < \frac{\varepsilon}{2}.\end{aligned} \tag{15}$$

Define now a reservoir system of the type (9) with inputs given by the random variables $\mathbf{Z}_t$, $t \in \mathbb{Z}_-$ and reservoir matrices $A \in \mathbb{M}_N$ and $c \in \mathbb{M}_{N,n}$ with all entries equal to 0 except $A_{i,i-n} = 1$ for $i = n+1, \ldots, N$ and $c_{i,i} = 1$ for $i = 1, \ldots, n$, that is

$$A = \begin{pmatrix} \mathbf{0}_{n,nK} & \mathbf{0}_{n,n} \\ \boldsymbol{I}_{nK} & \mathbf{0}_{n,n} \end{pmatrix}, \quad \text{and} \quad c = \begin{pmatrix} \boldsymbol{I}_n \\ \mathbf{0}_{nK,n} \end{pmatrix}.$$

This system has the echo state property (all the eigenvalues of A equal zero) and has a unique causal and time invariant solution associated to the reservoir states $\mathbf{X}_t := \left(\mathbf{Z}_t^\top, \mathbf{Z}_{t-1}^\top, \ldots, \mathbf{Z}_{t-K}^\top\right)^\top$, $t \in \mathbb{Z}_-$. It is easy to verify that the corresponding reservoir functional is given by

$$H_h^{A,c}(\mathbf{Z}) = h(\mathbf{Z}_0^\top, \ldots, \mathbf{Z}_{-K}^\top). \tag{16}$$

Now the triangle inequality and (14), (15) and (16) allow us to conclude (11).

The statement in (12) in the presence of the stationarity hypothesis for $\mathbf{Z}$ is a straightforward consequence of (6) and the equality (7). □

Remark III.2. A sufficient condition for (10) to hold is that the random variables $\{\mathbf{Z}_t : t \in \mathbb{Z}_-\}$ are independent and that for each t, there exists a constant $\alpha > 0$ such that $\mathbb{E}[\exp(\alpha \sum_{i=1}^n |Z_t^{(i)}|)] < \infty$.

Remark III.3. Assumption (10) can be replaced by alternative assumptions but it can not be removed. Even if $n = 1$ and $\{\mathbf{Z}_t : t \in \mathbb{Z}_-\}$ are independent and identically distributed with distribution ν, a condition *stronger* than the existence of moments of all orders for ν is required. As a counterexample, one may take for ν a lognormal distribution. Then ν has moments of all orders, but (10) is not satisfied. Let us now argue that the approximation result proved under assumption (10) fails in this case. The following argument relies on results for the classical *moment problem* (see, for example, the collection of references in [25]).

Indeed, by [26] ν is not determinate (there exist other probability measures with identical moments) and thus (see e.g. [27, Theorem 4.3]) Pol_1 is not dense in $L^p(\mathbb{R}, \nu)$ for $p \geq 2$. In particular, there exists $g \in L^p(\mathbb{R}, \nu)$ and $\varepsilon > 0$ such that $\|g - \tilde{h}\|_p > \varepsilon$ for all $\tilde{h} \in \mathrm{Pol}_1$. Suppose that we are in the case $n = 1$ and let $\{Z_t : t \in \mathbb{Z}_-\}$ be independent and identically distributed with distribution ν and $H(\mathbf{z}) := g(z_0)$ for $\mathbf{z} \in \mathbb{R}^{\mathbb{Z}_-}$. Then, for any choice of N, A, c and h one has $\mathbb{E}[H_h^{A,c}(\mathbf{Z})|\mathcal{F}_0] = \tilde{h}(Z_0)$, where $\tilde{h}(x) := \mathbb{E}[h(A\mathbf{X}_{-1} + cx)]$, $x \in \mathbb{R}$, is a polynomial. Thus one may use [24, Theorem 5.1.4] and the fact that by construction $H(\mathbf{Z})$ is $\mathcal{F}_0$-measurable to obtain

$$\begin{aligned} \|H(\mathbf{Z}) - H_h^{A,c}(\mathbf{Z})\|_p &\geq \|\mathbb{E}[H(\mathbf{Z})|\mathcal{F}_0] - \mathbb{E}[H_h^{A,c}(\mathbf{Z})|\mathcal{F}_0]\|_p \\ &= \|g - \tilde{h}\|_p > \varepsilon. \end{aligned}$$

Remark III.4. In previous reservoir computing universality results for both deterministic and stochastic inputs quoted in the introduction there was an important continuity hypothesis called the fading memory property that does not play a role here and that has been replaced by the integrability requirement $H \in L^p((\mathbb{R}^n)^{\mathbb{Z}_-}, \mu_{\mathbf{Z}})$. In particular, the universality results that we just proved and those that come in the next section (see Theorem III.7) yield approximations for filters which do not necessarily have the fading memory property. Whether or not the approximation results apply depends on the integrability condition with respect to the input environment measure $\mu_{\mathbf{Z}}$. Consider, for example, the functional associated to the peak-hold operator [9]. In the discrete-time setting, the associated functional is

$$H(\mathbf{z}) = \sup_{t \leq 0} z_t, \quad \text{with} \quad \mathbf{z} \in \mathbb{R}^{\mathbb{Z}_-}.$$

We now show that the two possibilities $H \in L^p((\mathbb{R}^n)^{\mathbb{Z}_-}, \mu_{\mathbf{Z}})$ and $H \notin L^p((\mathbb{R}^n)^{\mathbb{Z}_-}, \mu_{\mathbf{Z}})$ are feasible, depending on the choice of $\mu_{\mathbf{Z}}$:

- Let $\mathbf{Z} = (Z_t)_{t \in \mathbb{Z}_-}$ be one dimensional independent and identically distributed (i.i.d) random variables with unbounded support and denote by $\mu_{\mathbf{Z}}$ the law of $\mathbf{Z}$ on $\mathbb{R}^{\mathbb{Z}_-}$. Denoting by F the distribution function of Z_1 and using the i.i.d assumption one calculates, for any $a \in \mathbb{R}$,
$$\begin{aligned} \mathbb{P}(H(\mathbf{Z}) > a) &= 1 - \mathbb{P}(\cap_{t<0}\{Z_t \leq a\}) \\ &= 1 - \lim_{n \to \infty} F(a)^n = 1. \end{aligned}$$
Hence, we can conclude that $H(\mathbf{Z}) = \infty$, $\mu_{\mathbf{Z}}$-almost everywhere and therefore $H \notin L^p((\mathbb{R}^n)^{\mathbb{Z}_-}, \mu_{\mathbf{Z}})$.
- Consider now the same setup, but assume this time that the random variables have bounded support, that is, for some $a_{\max} \in \mathbb{R}$ one has that $P(Z_t \leq a_{\max}) = 1$ and $P(Z_t > a_{\max}) = 0$. Then, the same argument shows that $H(\mathbf{Z}) = a_{\max}$, $\mu_{\mathbf{Z}}$-almost everywhere and therefore $H \in L^p((\mathbb{R}^n)^{\mathbb{Z}_-}, \mu_{\mathbf{Z}})$.

Remark III.5. From the proof of Proposition III.1 one sees that one could replace in its statement Pol_N by any other family $\{\mathcal{H}_N\}_{N \in \mathbb{N}}$ that satisfies the density statement (13). In particular, the following corollary shows that this result can be obtained with readouts made out of neural networks.

Denote by $\mathcal{H}_N$ the set feedforward one hidden layer neural networks with inputs in $\mathbb{R}^N$ that are constructed with a fixed activation function σ. More specifically, $\mathcal{H}_N$ is made of functions $h : \mathbb{R}^N \to \mathbb{R}$ of the type

$$h(\mathbf{x}) = \sum_{j=1}^k \beta_j \sigma(\boldsymbol{\alpha}_j \cdot \mathbf{x} - \theta_j), \tag{17}$$

for some $k \in \mathbb{N}$, $\beta_j, \theta_j \in \mathbb{R}$, and $\boldsymbol{\alpha}_j \in \mathbb{R}^N$, for $j = 1, \ldots, k$.

Corollary III.6. *In the setup of Proposition III.1, consider the family of neural networks $h \in \mathcal{H}_N$ constructed with a fixed activation function σ that is bounded and non-constant. Then, for any $\varepsilon > 0$ there exists $N \in \mathbb{N}$, $A \in \mathbb{M}_N$, $c \in \mathbb{M}_{N,n}$, and a neural network $h \in \mathcal{H}_N$ such that the corresponding reservoir system (8) has the echo state property and has a unique causal and time-invariant filter associated. Moreover, the functional $H_h^{A,c}(\mathbf{Z}) \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ and satisfies that*

$$\|H(\mathbf{Z}) - H_h^{A,c}(\mathbf{Z})\|_p < \varepsilon. \tag{18}$$

Proof. By [6, Theorem 1] the set $\mathcal{H}_N$ is dense in $L^p(\mathbb{R}^N, \mu)$ for any finite measure μ on $\mathbb{R}^N$. Thus, statement (13) holds with $\mathcal{H}_N$ replacing $\mathrm{Pol}_{n(K+1)}$. Mimicking line by line the proof of Step 2 in Proposition III.1 then proves the Corollary. □

B. Trigonometric state-affine systems with linear readouts

Fix $M, N \in \mathbb{N}$ and consider $R\colon \mathbb{R}^n \to \mathbb{M}_{N,M}$ defined by

$$R(\mathbf{z}) := \sum_{k=1}^{r} A_k \cos(\mathbf{u}_k \cdot \mathbf{z}) + B_k \sin(\mathbf{v}_k \cdot \mathbf{z}), \quad \mathbf{z} \in \mathbb{R}^n, \tag{19}$$

for some $r \in \mathbb{N}$, $A_k, B_k \in \mathbb{M}_{N,M}$, $\mathbf{u}_k, \mathbf{v}_k \in \mathbb{R}^n$, for $k = 1, \dots, r$. The symbol $\mathrm{Trig}_{N,M}$ denotes the set of all functions of the type (19). We call the elements of $\mathrm{Trig}_{N,M}$ trigonometric polynomials.

We now introduce reservoir systems with linear readouts and reservoir maps constructed using trigonometric polynomials: let $N \in \mathbb{N}$, $\mathbf{W} \in \mathbb{R}^N$, $P \in \mathrm{Trig}_{N,N}$, $Q \in \mathrm{Trig}_{N,1}$ and define, for any $\mathbf{z} \in (\mathbb{R}^n)^{\mathbb{Z}_-}$, the system:

$$\begin{cases} \mathbf{x}_t = P(\mathbf{z}_t)\mathbf{x}_{t-1} + Q(\mathbf{z}_t), \quad t \in \mathbb{Z}_-, \\ y_t = \mathbf{W}^\top \mathbf{x}_t, \quad t \in \mathbb{Z}_-. \end{cases} \tag{20}$$

We call the systems of this type trigonometric state-affine systems. When such a system has the echo state property and a unique causal and time-invariant solution for any input, we denote by $U^{P,Q}_{\mathbf{W}}$ the corresponding filter and by $H^{P,Q}_{\mathbf{W}}(\mathbf{z}) := y_0$ the associated functional. As in the previous section, we fix $p \in [1, \infty)$, $\mathbf{Z}$ an input process, and a functional H such that $H(\mathbf{Z}) \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ and we are interested in approximating $H(\mathbf{Z})$ by systems of the form $H^{P,Q}_{\mathbf{W}}(\mathbf{Z})$. Again, we will write $H^{P,Q}_{\mathbf{W}}(\mathbf{Z}) = Y_0$, where Y_0 is uniquely determined by the reservoir system with stochastic inputs

$$\begin{cases} \mathbf{X}_t = P(Z_t)\mathbf{X}_{t-1} + Q(Z_t), \quad t \in \mathbb{Z}_-, \\ Y_t = \mathbf{W}^\top \mathbf{X}_t, \quad t \in \mathbb{Z}_-. \end{cases} \tag{21}$$

Define $\mathcal{A}$ as the set of four-tuples $(N, \mathbf{W}, P, Q) \in \mathbb{N} \times \mathbb{R}^N \times \mathrm{Trig}_{N,N} \times \mathrm{Trig}_{N,1}$ whose associated systems (20) have the echo state property and the unique solutions are causal and time-invariant. In particular, for such $(N, \mathbf{W}, P, Q)$ a reservoir functional $H^{P,Q}_{\mathbf{W}}$ associated to (20) exists.

Theorem III.7. *Let $p \in [1, \infty)$ and let $\mathbf{Z}$ be a fixed $\mathbb{R}^n$-valued input process. Denote by $\mathcal{L}_{\mathbf{Z}}$ the set of reservoir functionals of the type (20) which are p-integrable, that is,*

$$\mathcal{L}_{\mathbf{Z}} := \{H^{P,Q}_{\mathbf{W}}(\mathbf{Z}) \,:\, (N, \mathbf{W}, P, Q) \in \mathcal{A}\} \cap L^p(\Omega, \mathcal{F}, \mathbb{P}).$$

Then $\mathcal{L}_{\mathbf{Z}}$ is dense in $L^p(\Omega, \mathcal{F}_{-\infty}, \mathbb{P})$.

In particular, for any functional H such that $H(\mathbf{Z}) \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ and any $\varepsilon > 0$, there exists $N \in \mathbb{N}$, $\mathbf{W} \in \mathbb{R}^N$, $P \in \mathrm{Trig}_{N,N}$ and $Q \in \mathrm{Trig}_{N,1}$ such that the system (20) has the echo state property and causal and time-invariant solutions. Moreover, $H^{P,Q}_{\mathbf{W}}(\mathbf{Z}) \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ and

$$\|H(\mathbf{Z}) - H^{P,Q}_{\mathbf{W}}(\mathbf{Z})\|_p < \varepsilon. \tag{22}$$

If the input process $\mathbf{Z}$ is stationary then

$$\|U_H(\mathbf{Z}) - U^{P,Q}_{\mathbf{W}}(\mathbf{Z})\|_p < \varepsilon. \tag{23}$$

Proof. We first argue that $\mathcal{L}_{\mathbf{Z}}$ is a linear subspace of $L^p(\Omega, \mathcal{F}_{-\infty}, \mathbb{P})$. To do this we need to introduce some notation. Given $A \in \mathbb{M}_{N_1,M_1}$, $B \in M_{N_2,M_2}$, we denote by $A \oplus B \in \mathbb{M}_{N_1+N_2,M_1+M_2}$ the direct sum. Given R as in (19) we define $R \oplus A \in \mathrm{Trig}_{N+N_1,M+M_1}$ by

$$R \oplus A(\mathbf{z}) := \sum_{k=1}^{r} A_k \oplus A \cos(\mathbf{u}_k \cdot \mathbf{z}) + B_k \oplus A \sin(\mathbf{v}_k \cdot \mathbf{z}),$$

and (with the analogous definition for $B \oplus R$) for $R_i \in \mathrm{Trig}_{N_i,M_i}$, $i = 1, 2$ we set

$$R_1 \oplus R_2 = R_1 \oplus \mathbf{0}_{N_2,M_2} + \mathbf{0}_{N_1,M_1} \oplus R_2.$$

One easily verifies that for $\lambda \in \mathbb{R}$ and $(N_i, \mathbf{W}_i, P_i, Q_i) \in \mathcal{A}$, $i = 1, 2$, one has that

$$(N_1 + N_2, \mathbf{W}_1 \oplus \lambda \mathbf{W}_2, P_1 \oplus P_2, Q_1 \oplus Q_2) \in \mathcal{A},$$
$$H^{P_1,Q_1}_{\mathbf{W}_1}(\mathbf{Z}) + \lambda H^{P_2,Q_2}_{\mathbf{W}_2}(\mathbf{Z}) = H^{P_1 \oplus P_2, Q_1 \oplus Q_2}_{\mathbf{W}_1 \oplus \lambda \mathbf{W}_2}(\mathbf{Z}).$$

This shows that $\mathcal{L}_{\mathbf{Z}}$ is indeed a linear subspace of $L^p(\Omega, \mathcal{F}_{-\infty}, \mathbb{P})$.

Secondly, in order to show that $\mathcal{L}_{\mathbf{Z}}$ is dense in $L^p(\Omega, \mathcal{F}_{-\infty}, \mathbb{P})$, it suffices to prove that if $F \in L^q(\Omega, \mathcal{F}_{-\infty}, \mathbb{P})$ satisfies $\mathbb{E}[FH] = 0$ for all $H \in \mathcal{L}_{\mathbf{Z}}$, then $F = 0$, $\mathbb{P}$-almost surely. Here $q \in (1, \infty]$ is the Hölder conjugate exponent of p. This can be shown by contraposition. Suppose that $\mathcal{L}_{\mathbf{Z}}$ is not dense in $L^p(\Omega, \mathcal{F}_{-\infty}, \mathbb{P})$. Since $\mathcal{L}_{\mathbf{Z}}$ is a linear subspace, by the Hahn-Banach theorem there exists a bounded linear functional Λ on $L^p(\Omega, \mathcal{F}_{-\infty}, \mathbb{P})$ such that $\Lambda(H) = 0$ for all $H \in \mathcal{L}_{\mathbf{Z}}$, but $\Lambda \neq 0$, see e.g. [28, Theorem 5.19]. Then by [28, Theorem 6.16] there exists $F \in L^q(\Omega, \mathcal{F}_{-\infty}, \mathbb{P})$ such that $\Lambda(H) = \mathbb{E}[FH]$ for all $H \in L^p(\Omega, \mathcal{F}_{-\infty}, \mathbb{P})$ and $F \neq 0$, since $\Lambda \neq 0$. In particular, there exists $F \in L^q(\Omega, \mathcal{F}_{-\infty}, \mathbb{P}) \setminus \{0\}$ such that $\mathbb{E}[FH] = 0$ for all $H \in \mathcal{L}_{\mathbf{Z}}$.

Thirdly, suppose that $F \in L^q(\Omega, \mathcal{F}_{-\infty}, \mathbb{P})$ satisfies

$$\mathbb{E}[FH] = 0 \text{ for all } H \in \mathcal{L}_{\mathbf{Z}}. \tag{24}$$

If we show that $F = 0$, $\mathbb{P}$-almost surely, then the statement in the theorem follows by the argument in the second step. In order to prove that $F = 0$, $\mathbb{P}$-almost surely, we first show that (24) implies the following statement: for any $K \in \mathbb{N}$, any subset $I \subset \mathcal{I}_K := \{0, \dots, K\}$, and any $\mathbf{u}_0, \dots, \mathbf{u}_K \in \mathbb{R}^n$ it holds that

$$\mathbb{E}\left[F \prod_{j \in I} \sin(\mathbf{u}_j \cdot \mathbf{Z}_j) \prod_{k \in \mathcal{I}_K \setminus I} \cos(\mathbf{u}_k \cdot \mathbf{Z}_k)\right] = 0. \tag{25}$$

We prove this claim by induction on $K \in \mathbb{N}$. For $K = 0$, one sets $Q_1(\mathbf{z}) := \cos(\mathbf{u}_0 \cdot \mathbf{z})$ and $Q_2(\mathbf{z}) := \sin(\mathbf{u}_0 \cdot \mathbf{z})$ and notices that $(1, 1, 0, Q_i) \in \mathcal{A}$. Moreover, since the sine and cosine function are bounded, it is easy see that $Q_i(\mathbf{Z}_0) = H^{0,Q_i}_1(\mathbf{Z}_0) \in \mathcal{L}_{\mathbf{Z}}$, for $i \in \{1, 2\}$. Thus (24) implies (25) and so the statement holds for $K = 0$. For the induction step, let $K \in \mathbb{N} \setminus \{0\}$ and assume the implication holds for $K - 1$. We now fix I and $\mathbf{u}_0, \dots, \mathbf{u}_K \in \mathbb{R}^n$ as above and prove (25). To simplify the notation we define for $k \in \{0, \dots, K\}$ and $\mathbf{z} \in \mathbb{R}^n$ the function g_k by

$$g_k(\mathbf{z}) := \begin{cases} \sin(\mathbf{u}_k \cdot \mathbf{z}), & \text{if } k \in I, \\ \cos(\mathbf{u}_k \cdot \mathbf{z}), & \text{if } k \in \mathcal{I}_K \setminus I. \end{cases}$$

To prove (25), we set $N := K+1$, for $j \in \{1,\dots,K\}$ define $A_j \in \mathbb{M}_N$ with all entries equal to 0 except $(A_j)_{j+1,j} = 1$, that is, $(A_j)_{k,l} = \delta_{k,j+1}\delta_{l,j}$, $k,l \in \{1,\dots,N\}$. Define now for $\mathbf{z} \in \mathbb{R}^n$

$$\begin{cases} P(\mathbf{z}) := \sum_{j=0}^{K-1} A_{K-j} g_j(\mathbf{z}), \\ Q(\mathbf{z}) := e_1 g_K(\mathbf{z}), \\ \mathbf{W} := e_{K+1}, \end{cases} \tag{26}$$

where e_j is the j-th unit vector in $\mathbb{R}^N$, that is, the only non-zero entry of e_j is a 1 in the j-th coordinate. By Lemma A.2 in the appendix, one has $A_{j_L}\cdots A_{j_0} = 0$ for any $j_0,\dots,j_L \in \{1,\dots,K\}$ and $L \geq K$, since $j_L = j_0 + L$ can not be satisfied. In other words, any product of more than K factors of matrices $A^{(j)}$ is equal to 0 and thus for any $L \in \mathbb{N}$ with $L \geq K$ and any $\mathbf{z}_0,\dots,\mathbf{z}_L \in \mathbb{R}^n$ one has $P(\mathbf{z}_0)\dots P(\mathbf{z}_L) = 0$. Using this fact and iterating (20), one obtains that the trigonometric state-affine system defined by the elements in (26) has a unique solution given by

$$\mathbf{x}_t = Q(\mathbf{z}_t) + \sum_{j=1}^{K} P(\mathbf{z}_t)\cdots P(\mathbf{z}_{t-j+1})Q(\mathbf{z}_{t-j}). \tag{27}$$

In particular $(N,\mathbf{W},P,Q) \in \mathcal{A}$ and

$$\begin{aligned} &H^{P,Q}_{\mathbf{W}}(\mathbf{Z}) = \mathbf{X}_0 \\ &= \mathbf{W}^\top \left(Q(\mathbf{Z}_0) + \sum_{j=1}^{K} P(\mathbf{Z}_0)\cdots P(\mathbf{Z}_{-j+1})Q(\mathbf{Z}_{-j}) \right). \end{aligned} \tag{28}$$

The finiteness of the sum in (28) and the boundedness of the trigonometric polynomials implies that $H^{P,Q}_{\mathbf{W}}(\mathbf{Z}) \in \mathcal{L}_{\mathbf{Z}}$.

We conclude the proof of the induction step with the following chain of equalities that uses (24) in the first one, the representation (28) in the second one, and the choice of the vector $\mathbf{W}$ and the induction hypothesis in the last step:

$$\begin{aligned} 0 &= \mathbb{E}[F H^{P,Q}_{\mathbf{W}}(\mathbf{Z})] \\ &= \mathbb{E}[F\mathbf{W}^\top Q(\mathbf{Z}_0)] \\ &\quad + \mathbb{E}[F\mathbf{W}^\top \sum_{j=1}^{K} P(\mathbf{Z}_0)\cdots P(\mathbf{Z}_{-j+1})Q(\mathbf{Z}_{-j})] \\ &= \mathbb{E}[F\mathbf{W}^\top P(\mathbf{Z}_0)\cdots P(\mathbf{Z}_{-K+1})Q(\mathbf{Z}_{-K})]. \end{aligned} \tag{29}$$

However, again by Lemma A.2 in the appendix, the only non-zero product of matrices $A_{j_{K-1}}\cdots A_{j_0}$ for $j_0,\dots j_{K-1} \in \{1,\dots,K\}$ takes place when $j_k = k+1$ for $k \in \{0,\dots,K-1\}$. Therefore:

$$\begin{aligned} &P(\mathbf{Z}_0)\cdots P(\mathbf{Z}_{-K+1}) \\ &= A_K g_0(\mathbf{Z}_0) A_{K-1} g_1(\mathbf{Z}_{-1})\cdots A_1 g_{K-1}(\mathbf{Z}_{-K+1}). \end{aligned}$$

Combining this with (29) and using the identity (48) in Lemma A.2 in the appendix one obtains

$$\begin{aligned} 0 &= \mathbb{E}[F e_{K+1}^\top A_K \cdots A_1 e_1 \prod_{k=0}^{K} g_k(\mathbf{Z}_{-k})] \\ &= \mathbb{E}[F \prod_{k=0}^{K} g_k(\mathbf{Z}_{-k})], \end{aligned}$$

which is the same as (25).

Fourthly, by standard trigonometric identities, the identity (25) established in the third step implies that for any $K \in \mathbb{N}$,

$$\mathbb{E}\left[F \exp\left(i \sum_{j=0}^{K} \mathbf{u}_j \cdot \mathbf{Z}_j \right) \right] = 0 \text{ for all } \mathbf{u}_0,\dots,\mathbf{u}_K \in \mathbb{R}^n. \tag{30}$$

We claim that (30) implies $F = 0$, $\mathbb{P}$-almost surely and hence the statement in the theorem follows. This fact is a consequence of the *uniqueness* theorem for characteristic functions (which is ultimately a consequence of the Stone-Weierstrass approximation theorem). See for instance [21, Theorem 4.3] and the text below that result. To prove $F = 0$, $\mathbb{P}$-almost surely, we denote by F^+ and F^- the positive and negative parts of F. Then by (30) one has $\mathbb{E}[F] = 0$, necessarily. Thus, if it does not hold that $F = 0$, $\mathbb{P}$-almost surely, then $c := \mathbb{E}[F^+] = \mathbb{E}[F^-] > 0$ and one may define probability measures $\mathbb{Q}^+$ and $\mathbb{Q}^-$ on $(\Omega,\mathcal{F})$ by setting $\mathbb{Q}^+(A) := c^{-1}\mathbb{E}[F^+\mathbb{1}_A]$ and $\mathbb{Q}^-(A) := c^{-1}\mathbb{E}[F^-\mathbb{1}_A]$ for $A \in \mathcal{F}$. Denote by μ_K^+ and μ_K^- the law in $\mathbb{R}^{n(K+1)}$ of the random variable

$$\boldsymbol{\mathcal{Z}}_K := (\mathbf{Z}_0^\top, \mathbf{Z}_{-1}^\top, \dots, \mathbf{Z}_{-K}^\top)^\top$$

under $\mathbb{Q}^+$ and $\mathbb{Q}^-$. Then, the statement (30) implies that for all $u \in \mathbb{R}^{n(K+1)}$,

$$\int_{\mathbb{R}^{n(K+1)}} \exp(iu\cdot z)\mu_K^+(\mathrm{d}z) = \int_{\mathbb{R}^{n(K+1)}} \exp(iu\cdot \mathbf{z})\mu_K^-(\mathrm{d}\mathbf{z}).$$

By the uniqueness theorem for characteristic functions (see e.g. [21, Theorem 4.3] and the text below) this implies that $\mu_K^+ = \mu_K^-$. Translating this statement back to random variables, this means that for any bounded and measurable function $g\colon \mathbb{R}^{n(K+1)} \to \mathbb{R}$ one has

$$0 = c\mathbb{E}_{\mathbb{Q}^+}[g(\boldsymbol{\mathcal{Z}}_K)] - c\mathbb{E}_{\mathbb{Q}^-}[g(\boldsymbol{\mathcal{Z}}_K)] = \mathbb{E}[F g(\boldsymbol{\mathcal{Z}}_K)],$$

which, by definition, means that $\mathbb{E}[F|\mathcal{F}_{-K}] = 0$, $\mathbb{P}$-almost surely. Since $K \in \mathbb{N}$ was arbitrary and $F \in L^1(\Omega,\mathcal{F}_{-\infty},\mathbb{P})$, one may combine this with $\lim_{t\to-\infty}\mathbb{E}[F|\mathcal{F}_t] = F$, $\mathbb{P}$-almost surely (see Lemma A.1) to conclude $F = 0$, as desired.

The statement in (23) in the presence of the stationarity hypothesis for $\mathbf{Z}$ is a straightforward consequence of (6) and the equality (7). □

We emphasize that the use in the proof of the theorem of nilpotent matrices of the type introduced in Lemma A.2 ensures that the the echo state property is automatically satisfied (see (27)).

C. Echo state networks

We now turn to showing the universality in the L^p sense of the the most widely used reservoir systems with linear readouts, namely, echo state networks. An echo state network is a RC system determined by

$$\begin{cases} \mathbf{x}_t = \sigma(A\mathbf{x}_{t-1} + C\mathbf{z}_t + \boldsymbol{\zeta}), \\ y_t = \mathbf{W}^\top \mathbf{x}_t, \end{cases} \tag{31}$$

for $A \in \mathbb{M}_N$, $C \in M_{N,n}$, $\boldsymbol{\zeta} \in \mathbb{R}^N$, and $\mathbf{W} \in \mathbb{R}^N$. As it is customary in the neural networks literature, the map σ :

$\mathbb{R}^N \to \mathbb{R}^N$ is obtained via the componentwise application of a given activation function $\sigma\colon \mathbb{R} \to \mathbb{R}$ that is denoted with the same symbol.

If this system has the echo state property and the resulting filter is causal and time-invariant, we write as $H^{A,C,\boldsymbol{\zeta}}_{\mathbf{W}}(\mathbf{z}) := y_0$ the associated functional.

Theorem III.8. *Fix $p \in [1,\infty)$, let $\mathbf{Z}$ be a fixed $\mathbb{R}^n$-valued input process, and let H be a functional such that $H(\mathbf{Z}) \in L^p(\Omega,\mathcal{F},\mathbb{P})$. Suppose that the activation function $\sigma\colon \mathbb{R} \to \mathbb{R}$ is non-constant, continuous, and has a bounded image. Then for any $\varepsilon > 0$, there exists $N \in \mathbb{N}$, $C \in \mathbb{M}_{N,n}$, $\boldsymbol{\zeta} \in \mathbb{R}^N$, $A \in \mathbb{M}_N$, $\mathbf{W} \in \mathbb{R}^N$ such that (31) has the echo state property, the corresponding filter is causal and time-invariant, the associated functional satisfies $H^{A,C,\boldsymbol{\zeta}}_{\mathbf{W}}(\mathbf{Z}) \in L^p(\Omega,\mathcal{F},\mathbb{P})$ and*

$$\|H(\mathbf{Z}) - H^{A,C,\boldsymbol{\zeta}}_{\mathbf{W}}(\mathbf{Z})\|_p < \varepsilon. \tag{32}$$

Proof. First, by Corollary III.6 and (16) there exists $K, \overline{N} \in \mathbb{N}$, $\overline{\mathbf{W}} \in \mathbb{R}^{\overline{N}}$, $\overline{A} \in \mathbb{M}_{\overline{N},n(K+1)}$, and $\overline{\boldsymbol{\zeta}} \in \mathbb{R}^{\overline{N}}$ such that the neural network

$$h(\mathbf{z}) = \overline{\mathbf{W}}^\top \sigma(\overline{A}\mathbf{z} + \overline{\boldsymbol{\zeta}})$$

satisfies

$$\|H(\mathbf{Z}) - h(\mathbf{Z}_0^\top, \dots, \mathbf{Z}_{-K}^\top)\|_p < \frac{\varepsilon}{2}. \tag{33}$$

Notice that we may rewrite $\overline{A}$ as

$$\overline{A} = [A^{(0)} A^{(-1)} \cdots A^{(-K)}]$$

with $A^{(j)} \in \mathbb{M}_{\overline{N},n}$ and

$$\begin{aligned} H_\infty(\mathbf{Z}) :&= h(\mathbf{Z}_0^\top, \dots, \mathbf{Z}_{-K}^\top) \\ &= \overline{\mathbf{W}}^\top \sigma\left(\sum_{j=0}^{K} A^{(-j)}\mathbf{Z}_{-j} + \overline{\boldsymbol{\zeta}}\right). \end{aligned} \tag{34}$$

Second, by the neural network approximation theorem for continuous functions [6, Theorem 2], for any $m \in \mathbb{N}$ there exists a neural network that uniformly approximates the identity mapping on the hypercube $B_m := \{\mathbf{x} \in \mathbb{R}^n : |x_i| \le m \text{ for } i = 1,\dots,n\}$. More specifically, [6, Theorem 2] is formulated for $\mathbb{R}$-valued mappings and we hence apply it componentwise: for any $m \in \mathbb{N}$ and $i = 1,\dots,n$ there exists $N_i^{(m)} \in \mathbb{N}$, $\mathbf{W}_i^{(m)} \in \mathbb{R}^{N_i^{(m)}}$, $\overline{A}_i^{(m)} \in \mathbb{M}_{N_i^{(m)},n}$, and $\boldsymbol{\zeta}_i^{(m)} \in \mathbb{R}^{N_i^{(m)}}$, such that for all $i = 1,\dots,n$ the neural network

$$h_i^{(m)}(\mathbf{x}) = \left(\mathbf{W}_i^{(m)}\right)^\top \sigma\left(\overline{A}_i^{(m)}\mathbf{x} + \boldsymbol{\zeta}_i^{(m)}\right)$$

satisfies

$$\sup_{\mathbf{x}\in B_m} |h_i^{(m)}(\mathbf{x}) - x_i| < \frac{1}{m}. \tag{35}$$

Write $h^{(m)}(\mathbf{x}) = (h_1^{(m)}(\mathbf{x}),\dots,h_n^{(m)}(\mathbf{x}))^\top$ and for $j = 1,\dots,K$, denote by $[h^{(m)}]^j = h^{(m)} \circ \cdots \circ h^{(m)}$ the jth composition of $h^{(m)}$. We now claim that for all $j = 1,\dots,K$ and $\mathbf{x} \in \mathbb{R}^n$ it holds that

$$\lim_{m\to\infty} [h^{(m)}]^j(\mathbf{x}) = \mathbf{x}. \tag{36}$$

Indeed, let us fix $\mathbf{x} \in \mathbb{R}^n$ and argue by induction on j. To prove (36) for $j = 1$, let $\overline{\varepsilon} > 0$ be given and choose $m_0 \in \mathbb{N}$ satisfying $m_0 > \max\{|x_1|,\dots,|x_n|, 1/\overline{\varepsilon}\}$. Then, for any $m \ge m_0$ one has $\mathbf{x} \in B_m$ by definition and (35) implies that for $i = 1,\dots,n$,

$$|h_i^{(m)}(\mathbf{x}) - x_i| < \frac{1}{m} < \overline{\varepsilon}.$$

Hence (36) indeed holds for $j = 1$. Now let $j \ge 2$ and assume that (36) has been proved for $j-1$. Define $\overline{\mathbf{x}}^{(m)} := [h^{(m)}]^{j-1}(\mathbf{x})$. Then, by the induction hypothesis, for any given $\overline{\varepsilon} > 0$ one finds $m_0 \in \mathbb{N}$ such that for all $m \ge m_0$ and $i = 1,\dots,n$ it holds that

$$|\overline{x}_i^{(m)} - x_i| < \frac{\overline{\varepsilon}}{2}. \tag{37}$$

Hence, choosing $\overline{m}_0 \in \mathbb{N}$ with $\overline{m}_0 > \max(m_0, |x_1| + \frac{\overline{\varepsilon}}{2}, \dots, |x_n| + \frac{\overline{\varepsilon}}{2}, 2/\overline{\varepsilon})$ one obtains from the triangle inequality and (37) that $\overline{\mathbf{x}}^{(m)} \in B_{\overline{m}_0}$ for all $m \ge \overline{m}_0$. In particular for any $m \ge \overline{m}_0$ one may use the triangle inequality in the first step, $\overline{\mathbf{x}}^{(m)} \in B_{\overline{m}_0} \subset B_m$ and (37) in the second step and (35) in the last step to estimate

$$\begin{aligned} |[h^{(m)}]_i^j(\mathbf{x}) - x_i| &\le |h_i^{(m)}(\overline{\mathbf{x}}^{(m)}) - \overline{x}_i^{(m)}| + |\overline{x}_i^{(m)} - x_i| \\ &\le \sup_{\mathbf{y}\in B_m} |h_i^{(m)}(\mathbf{y}) - y_i| + \frac{\overline{\varepsilon}}{2} \\ &< \frac{1}{m} + \frac{\overline{\varepsilon}}{2} < \overline{\varepsilon}. \end{aligned}$$

This proves (36) for all $j = 1,\dots,K$.

Thirdly, define

$$H_m(\mathbf{Z}) := \overline{\mathbf{W}}^\top \sigma\left(\sum_{j=0}^{K} A^{(-j)}[h^{(m)}]^j(\mathbf{Z}_{-j}) + \overline{\boldsymbol{\zeta}}\right)$$

with the convention $[h^{(m)}]^0(\mathbf{x}) = \mathbf{x}$.

Since σ is continuous, (36) implies that $\lim_{m\to\infty} H_m(\mathbf{Z}) = H_\infty(\mathbf{Z})$, $\mathbb{P}$-almost surely, where H_∞ was defined in (34). Furthermore, by assumption there exists $C > 0$ such that $|\sigma(x)| \le C$ for all $x \in \mathbb{R}$. Hence one has $|H_\infty(\mathbf{Z}) - H_m(\mathbf{Z})|^p \le (2C\sum_{i=1}^{\overline{N}} |\overline{W}_i|)^p$ for all $m \in \mathbb{N}$. Thus one may apply the dominated convergence theorem to obtain

$$\begin{aligned} &\lim_{m\to\infty} \|H_\infty(\mathbf{Z}) - H_m(\mathbf{Z})\|_p \\ &= \lim_{m\to\infty} \mathbb{E}[|H_\infty(\mathbf{Z}) - H_m(\mathbf{Z})|^p]^{1/p} = 0. \end{aligned}$$

In particular for $m \in \mathbb{N}$ large enough one has $\|H_\infty(\mathbf{Z}) - H_m(\mathbf{Z})\|_p < \frac{\varepsilon}{2}$ and combining this with the triangle inequality and (33) one obtains

$$\begin{aligned} \|H(\mathbf{Z}) - H_m(\mathbf{Z})\|_p &\le \|H(\mathbf{Z}) - H_\infty(\mathbf{Z})\|_p \\ &\quad + \|H_\infty(\mathbf{Z}) - H_m(\mathbf{Z})\|_p < \varepsilon. \end{aligned} \tag{38}$$

To conclude the proof we now fix $m \in \mathbb{N}$ large enough (so that (38) holds) and show that $H_m(\mathbf{Z}) = H^{A,C,\boldsymbol{\zeta}}_{\mathbf{W}}(\mathbf{Z})$ for suitable choices of $A, C, \boldsymbol{\zeta}$ and $\mathbf{W}$. To do so, first define $N_J := N_1^{(m)} + \cdots + N_n^{(m)}$ and the block matrices

$$W_J := \begin{pmatrix} (\mathbf{W}_1^{(m)})^\top & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & (\mathbf{W}_n^{(m)})^\top \end{pmatrix} \in \mathbb{M}_{n,N_J},$$

$$\zeta_J := \begin{pmatrix} \zeta_1^{(m)} \\ \vdots \\ \zeta_n^{(m)} \end{pmatrix} \in \mathbb{R}^{N_J}, \text{ and } A_J := \begin{pmatrix} \overline{A}_1^{(m)} \\ \vdots \\ \overline{A}_n^{(m)} \end{pmatrix} \in \mathbb{M}_{N_J,n}.$$

Furthermore, to emphasize that m is fixed and $h^{(m)}$ approximates the identity, set $J(\mathbf{x}) := h^{(m)}(\mathbf{x})$ and note that

$$J(\mathbf{x}) = W_J \sigma(A_J \mathbf{x} + \zeta_J). \tag{39}$$

Now set $N := KN_J + \overline{N}$ and define the block matrix $A \in \mathbb{M}_N$ by

$$A = \begin{pmatrix} \mathbf{0}_{N_J,N_J} & & & & \\ A_J W_J & \mathbf{0}_{N_J,N_J} & & \mathbf{0} & \\ & A_J W_J & \ddots & & \\ \mathbf{0} & & \ddots & \mathbf{0}_{N_J,N_J} & \\ & & & A_J W_J & \\ A^{(-1)} W_J & A^{(-2)} W_J & \cdots & \cdots & A^{(-K)} W_J & \mathbf{0}_{\overline{N},\overline{N}} \end{pmatrix}$$

and $\zeta \in \mathbb{R}^N$, $C \in \mathbb{M}_{N,n}$ and $\mathbf{W} \in \mathbb{R}^N$ by

$$\zeta := \begin{pmatrix} \zeta_J \\ \vdots \\ \zeta_J \\ \overline{\zeta} \end{pmatrix}, \quad C := \begin{pmatrix} A_J \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ A^{(0)} \end{pmatrix}, \text{ and } \mathbf{W} := \begin{pmatrix} \mathbf{0}_{KN_J,1} \\ \overline{\mathbf{W}} \end{pmatrix}.$$

Furthermore, we partition the reservoir states $\mathbf{x}_t$ of the corresponding echo state system as

$$\mathbf{x}_t := \begin{pmatrix} \overline{\mathbf{x}}_t^{(1)} \\ \vdots \\ \overline{\mathbf{x}}_t^{(K+1)} \end{pmatrix},$$

with $\overline{\mathbf{x}}_t^{(j)} \in \mathbb{R}^{N_J}$, for $j \leq K$, and $\overline{\mathbf{x}}_t^{(K+1)} \in \mathbb{R}^{\overline{N}}$. With this notation for $\mathbf{x}_t$ and these choices of matrices, the recursions associated to the echo state reservoir map in (31) read as

$$\overline{\mathbf{x}}_t^{(1)} = \sigma(A_J \mathbf{z}_t + \zeta_J), \tag{40}$$

$$\overline{\mathbf{x}}_t^{(j)} = \sigma(A_J W_J \overline{\mathbf{x}}_{t-1}^{(j-1)} + \zeta_J), \text{ for } j = 2, \ldots, K, \tag{41}$$

$$\overline{\mathbf{x}}_t^{(K+1)} = \sigma(\sum_{j=1}^{K} A^{(-j)} W_J \overline{\mathbf{x}}_{t-1}^{(j)} + A^{(0)} \mathbf{z}_t + \overline{\zeta}). \tag{42}$$

By iteratively inserting (41) into itself and using (40) one obtains (recall the definition of J in (39)) that the unique solution to (41) is given by

$$\overline{\mathbf{x}}_t^{(j)} = \sigma(A_J [J]^{j-1}(\mathbf{z}_{t-j+1}) + \zeta_J). \tag{43}$$

More formally, one uses induction on j: For $j = 1$ the two expressions (43) and (40) coincide. For $j = 2, \ldots, K$ one inserts (43) for $j-1$ (which holds by induction hypothesis) into (41) to obtain

$$\begin{aligned} \overline{\mathbf{x}}_t^{(j)} &= \sigma(A_J W_J \sigma(A_J [J]^{j-2}(\mathbf{z}_{t-j+1}) + \zeta_J) + \zeta_J) \\ &= \sigma(A_J [J]^{j-1}(\mathbf{z}_{t-j+1}) + \zeta_J), \end{aligned}$$

which is indeed (43). Finally, combining (43) and (42) one obtains

$$\begin{aligned} y_t = \overline{\mathbf{W}}^\top \overline{\mathbf{x}}_t^{(K+1)} &= \overline{\mathbf{W}}^\top \sigma(\sum_{j=1}^{K} A^{(-j)} W_J \overline{\mathbf{x}}_{t-1}^{(j)} + A^{(0)} \mathbf{z}_t + \overline{\zeta}) \\ &= \overline{\mathbf{W}}^\top \sigma(\sum_{j=1}^{K} A^{(-j)} [J]^j (\mathbf{z}_{t-j}) + A^{(0)} \mathbf{z}_t + \overline{\zeta}). \end{aligned}$$

The statement (43) shows, in particular, that the echo state network associated to A, C, ζ and $\mathbf{W}$ satisfies the echo state property. Moreover, inserting $t = 0$ in the previous equality and comparing with the definition of $H_m(\mathbf{Z})$ one sees that indeed $H_m(\mathbf{Z}) = H_{\mathbf{W}}^{A,C,\zeta}(\mathbf{Z})$. The approximation statement (32) therefore follows from (38). □

D. An alternative viewpoint

So far all the universality results have been formulated for functionals and filters with random inputs. Equivalently, we may formulate them as L^p-approximation results on the sequence space $(\mathbb{R}^n)^{\mathbb{Z}_-}$ endowed with any measure μ that makes p-integrable the filter that we want to approximate.

Theorem III.9. *Let $H\colon (\mathbb{R}^n)^{\mathbb{Z}_-} \to \mathbb{R}$ be a measurable functional. Then, for any probability measure μ on $(\mathbb{R}^n)^{\mathbb{Z}_-}$ with $H \in L^p((\mathbb{R}^n)^{\mathbb{Z}_-}, \mu)$ and any $\varepsilon > 0$ there exists a reservoir system that has the echo state property and such that the corresponding filter is causal and time-invariant, the associated functional H^{RC} satisfies that $H^{RC} \in L^p((\mathbb{R}^n)^{\mathbb{Z}_-}, \mu)$ and*

$$\|H - H^{RC}\|_{L^p((\mathbb{R}^n)^{\mathbb{Z}_-}, \mu)} < \varepsilon. \tag{44}$$

The reservoir functional H^{RC} may be chosen as coming from any of the following systems:

- *Linear reservoir with polynomial readout, that is, (8) for some $N \in \mathbb{N}$, $A \in \mathbb{M}_N$, $\mathbf{c} \in \mathbb{M}_{N,n}$, and a polynomial $h \in \mathrm{Pol}_N$, if the measure μ satisfies the following condition: for any $K \in \mathbb{N}$,*

$$\int_{(\mathbb{R}^n)^{\mathbb{Z}_-}} \exp\left(\alpha \sum_{k=0}^{K} \sum_{i=1}^{n} |z_{-k}^{(i)}|\right) \mu(\mathrm{d}z) < \infty.$$

- *Linear reservoir with neural network readout, that is, (8) for some $N \in \mathbb{N}$, $A \in \mathbb{M}_N$, $\mathbf{c} \in \mathbb{M}_{N,n}$, and a neural network $h \in \mathcal{H}_N$.*
- *Trigonometric state-affine system with linear readout, that is, (20) for some $N \in \mathbb{N}$, $\mathbf{W} \in \mathbb{R}^N$, $P \in \mathrm{Trig}_{N,N}$ and $Q \in \mathrm{Trig}_{N,1}$.*
- *Echo state network with linear readout, that is, (31) for some $N \in \mathbb{N}$, $C \in \mathbb{M}_{N,n}$, $\zeta \in \mathbb{R}^N$, $A \in \mathbb{M}_N$, $\mathbf{W} \in \mathbb{R}^N$, where we assume that $\sigma\colon \mathbb{R} \to \mathbb{R}$ employed in (31) is bounded, continuous and non-constant.*

Proof. Set $\Omega = (\mathbb{R}^n)^{\mathbb{Z}_-}$, $\mathcal{F} = \otimes_{t \in \mathbb{Z}_-} \mathcal{B}(\mathbb{R}^n)$, $\mathbb{P} = \mu$ and define $\mathbf{Z}_t(\mathbf{z}) := \mathbf{z}_t$ for all $\mathbf{z} \in \Omega$, $t \in \mathbb{Z}_-$. Then $\mathcal{F} = \sigma(\mathbf{Z}_t : t \in \mathbb{Z}_-) = \mathcal{F}_{-\infty}$ and $\mathbf{Z}$ is the identity mapping on $(\mathbb{R}^n)^{\mathbb{Z}_-}$. One may now apply Proposition III.1, Corollary III.6, Theorem III.7 and Theorem III.8 with this choice of probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and input process $\mathbf{Z}$. The statement of Theorem III.9 then precisely coincides with the

statement of Proposition III.1, Corollary III.6, Theorem III.7 and Theorem III.8, respectively. □

E. Approximation of stationary strong time series models

Most parametric time series models commonly used in financial, macroeconometrics, and forecasting applications are specified by relations of the type

$$\mathbf{X}_t = G\left(\mathbf{X}_{t-1}, \mathbf{Z}_t, \boldsymbol{\theta}\right), \tag{45}$$

where $\boldsymbol{\theta} \in \mathbb{R}^k$ are the parameters of the model and the vector $\mathbf{X}_t \in \mathbb{R}^N$ is built so that it contains in its components the time series of interest and that, at the same time, allows for a Markovian representation of the model as in (45). The model is driven by the innovations process $\mathbf{Z} = (\mathbf{Z}_t)_{t\in\mathbb{Z}} \in (\mathbb{R}^n)^{\mathbb{Z}}$. When the innovations are made out of independent and identically distributed random variables we say that the model is strong [29]. It is customary in the time series literature to impose constraints on the parameters vector $\boldsymbol{\theta}$ so that the relation (45) has a unique second-order stationary solution or, in the language of this paper, the system (45) satisfies the echo state property and the associated filter $U_G : (\mathbb{R}^n)^{\mathbb{Z}} \to \left(\mathbb{R}^N\right)^{\mathbb{Z}}$ satisfies that

$$\mathbb{E}\left[U_G(\mathbf{Z})_t\right] =: \boldsymbol{\mu} \text{ and } \mathbb{E}\left[U_G(\mathbf{Z})_t U_G(\mathbf{Z})_t^\top\right] =: \boldsymbol{\Sigma},\ t \in \mathbb{Z}_-, \tag{46}$$

with $\boldsymbol{\mu} \in \mathbb{R}^N$ and $\boldsymbol{\Sigma} \in \mathbb{M}_N$ constant. The Wold decomposition theorem [30, Theorem 5.7.1] shows that any such filter can be uniquely written as the sum of a linear and a deterministic process.

It is obvious that for strong models the stationarity condition (6) holds and that, moreover, the condition (46) implies that

$$\|U_G(\mathbf{Z})\|_2 = \sup_{t\in\mathbb{Z}_-}\left\{\mathbb{E}\left[|U(\mathbf{Z})_t|^2\right]^{1/2}\right\} = \operatorname{trace}\left(\boldsymbol{\Sigma}\right)^{1/2} < \infty. \tag{47}$$

This integrability condition guarantees that the approximation results in Proposition III.1, Corollary III.6, and Theorems III.7 and III.8 hold for second-order stationary strong time series models with $p = 2$. More specifically, the processes determined by this kind of models can be approximated in the L^2 sense by linear processes with polynomial or neural network readouts (when the condition in Remark III.2 is satisfied), by trigonometric state-affine systems with linear readouts, or by echo state networks.

Important families of models to which this approximation statement can be applied are, among many others, (see the references for the meaning of the acronyms) GARCH [31], [32], VEC [33], BEKK [34], CCC [35], DCC [36], [37], GDC [38], and ARSV [39], [40].

IV. Conclusion

We have shown the universality of three different families of reservoir computers with respect to the L^p norm associated to any given discrete-time semi-infinite input process.

On the one hand we proved that linear reservoir systems with either neural network or, if the input process satisfies the exponential moments condition (10), polynomial readout maps are universal.

On the other hand we showed that this hypothesis can be dropped by considering two different reservoir families with linear readouts, namely, trigonometric state-affine systems and echo state networks. The latter are the most widely used reservoir systems in applications. The linearity in the readouts is a key feature in supervised machine learning applications of these systems. It guarantees that they can be used in high-dimensional situations and in the presence of large datasets, since the training in that case is reduced to a linear regression.

We emphasize that, unlike existing results in the literature [19], [20] dealing with uniform universal approximation, the L^p criteria used in this paper allow to formulate universality statements that do not necessarily impose almost sure uniform boundedness on the inputs or the fading memory property on the filter that needs to be approximated.

Appendix

A. Auxiliary Lemmas

Lemma A.1. *Let $\mathbf{Z} : \mathbb{Z}\times\Omega \to \mathbb{R}^n$ be a stochastic process and let $\mathcal{F}_t := \sigma(\mathbf{Z}_0, \ldots, \mathbf{Z}_t)$, $t \in \mathbb{Z}_-$, and $\mathcal{F}_{-\infty} := \sigma(\mathbf{Z}_t : t \in \mathbb{Z}_-)\}$. Let $F \in L^p(\Omega, \mathcal{F}_{-\infty}, \mathbb{P})$. Then $\mathbb{E}[F|\mathcal{F}_t]$ converges to F as $t \to -\infty$, both $\mathbb{P}$-almost surely and in norm $\|\cdot\|_p$, for any $p \in [1, \infty)$.*

Proof. Since $\mathcal{F}_{-t} \subset \mathcal{F}_{-t-1} \subset \mathcal{F}_{-\infty}$, for all $t \in \mathbb{N}$, and $F \in L^p(\Omega, \mathcal{F}_{-\infty}, \mathbb{P}) \subset L^1(\Omega, \mathcal{F}_{-\infty}, \mathbb{P})$, one has by Lévy's Upward Theorem (see, for instance, [41, II.50.3] or [24, Theorem 5.5.7]) that $F_t := \mathbb{E}[F|\mathcal{F}_t]$ converges for $t \to -\infty$ to F in $\|\cdot\|_1$ and $\mathbb{P}$-almost surely. If $p = 1$ this already implies the claim. For $p > 1$ one has by standard properties of conditional expectations (see, for instance, [24, Theorem 5.1.4]) that $\sup_{t\in\mathbb{N}} \mathbb{E}[|F_t|^p] \leq \mathbb{E}[|F|^p]$. Hence [24, Theorem 5.4.5] implies that F_t converges for $t \to -\infty$ to some $\tilde{F} \in L^p(\Omega, \mathcal{F}_{-\infty}, \mathbb{P})$ both in $\|\cdot\|_p$ and $\mathbb{P}$-almost surely. But this identifies $\tilde{F} = \lim_{t\to-\infty} F_t = F$, $\mathbb{P}$-almost surely and hence F_t converges for $t \to -\infty$ to F also in $\|\cdot\|_p$. □

Lemma A.2. *For $N \in \mathbb{N}\setminus\{0,1\}$ and $j = 1, \ldots, N-1$ define $A_j \in \mathbb{M}_N$ by $(A_j)_{k,l} = \delta_{k,j+1}\delta_{l,j}$ for $k, l \in \{1, \ldots, N\}$. Then for $L \in \mathbb{N}$, $j_0, \ldots, j_L \in \{1, \ldots, N-1\}$ it holds that*

$$(A_{j_L} \cdots A_{j_0})_{k,l} = \delta_{k,j_L+1}\delta_{l,j_0}\prod_{i=1}^{L}\delta_{j_i,j_{i-1}+1}. \tag{48}$$

In particular $A_{j_L} \cdots A_{j_0} \neq 0$ if and only if $j_i = j_0 + i$ for $i \in \{1, \ldots, L\}$.

Proof. The last statement directly follows from (48). To prove (48) we proceed by induction on L. Indeed, for $L = 0$ the formula (48) is just the definition of A_{j_0}. For the induction step, one assumes that (48) holds for $L-1$ and calculates

$$\begin{aligned}(A_{j_L} \cdots A_{j_0})_{k,l} &= \sum_{r=1}^{N}\delta_{k,j_L+1}\delta_{r,j_L}(A_{j_{L-1}} \cdots A_{j_0})_{r,l}\\ &= \sum_{r=1}^{N}\delta_{k,j_L+1}\delta_{r,j_L}\delta_{r,j_{L-1}+1}\delta_{l,j_0}\prod_{i=1}^{L-1}\delta_{j_i,j_{i-1}+1},\end{aligned}$$

which is indeed (48). □

ACKNOWLEDGMENT

The authors thank Lyudmila Grigoryeva and Josef Teichmann for helpful discussions and remarks and acknowledge partial financial support coming from the Research Commission of the Universität Sankt Gallen, the Swiss National Science Foundation (grants number 175801/1 and 179114), and the French ANR "BIPHOPROC" project (ANR-14-OHRI-0018-02).

REFERENCES

[1] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bulletin of the American Mathematical Society*, vol. 39, no. 1, pp. 1–49, 2002.

[2] S. Smale and D.-X. Zhou, "Estimating the approximation error in learning theory," *Analysis and Applications*, vol. 01, no. 01, pp. 17–41, jan 2003. [Online]. Available: http://www.worldscientific.com/doi/abs/10.1142/S0219530503000089

[3] F. Cucker and D.-X. Zhou, *Learning Theory : An Approximation Theory Viewpoint*. Cambridge University Press, 2007. [Online]. Available: http://www.cambridge.org/fr/academic/subjects/computer-science/pattern-recognition-and-machine-learning/learning-theory-approximation-theory-viewpoint?format=HB{&}isbn=9780521865593{#}GuqpPouY1TsVIjpJ.97

[4] G. Cybenko, "Approximation by superpositions of a sigmoidal function," pp. 303–314, dec 1989.

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

[6] K. Hornik, "Approximation capabilities of muitilayer feedforward networks," *Neural Networks*, vol. 4, no. 1989, pp. 251–257, 1991.

[7] 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.

[8] 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.

[9] 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, 1985.

[10] 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. [Online]. Available: http://ieeexplore.ieee.org/document/1084646/

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

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

[13] I. W. Sandberg, "Approximation theorems for discrete-time systems," *IEEE Transactions on Circuits and Systems*, vol. 38, no. 5, pp. 564–566, 1991.

[14] ——, "Structure theorems for nonlinear systems," *Multidimensional Systems and Signal Processing*, vol. 2, pp. 267–286, 1991.

[15] M. B. Matthews, "On the Uniform Approximation of Nonlinear Discrete-Time Fading-Memory Systems Using Neural Network Models," Ph.D. dissertation, ETH Zürich, 1992. [Online]. Available: https://www.research-collection.ethz.ch:443/handle/20.500.11850/140592

[16] ——, "Approximating nonlinear fading-memory operators using neural network models," *Circuits, Systems, and Signal Processing*, vol. 12, no. 2, pp. 279–307, jun 1993.

[17] P. C. Perryman, "Approximation Theory for Deterministic and Stochastic Nonlinear Systems," Ph.D. dissertation, University of California, Irvine, 1996.

[18] 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*, vol. 1. IEEE Comput. Soc. Press, 1997, pp. 141–145.

[19] L. Grigoryeva and J.-P. Ortega, "Universal discrete-time reservoir computers with stochastic inputs and linear readouts using non-homogeneous state-affine systems," *Preprint*, dec 2017. [Online]. Available: http://arxiv.org/abs/1712.00754

[20] ——, "Echo state networks are universal," *Preprint*, 2018. [Online]. Available: https://arxiv.org/abs/1806.00797

[21] O. Kallenberg, *Foundations of Modern Probability*, ser. Probability and Its Applications. Springer New York, 2002.

[22] C. Berg and J. P. R. Christensen, "Density questions in the classical theory of moments," *Annales de l'Institut Fourier*, vol. 31, no. 3, pp. 99–114, 1981.

[23] L. C. Petersen, "On the relation between the multidimensional moment problem and the one-dimensional moment problem," *Mathematica Scandinavica*, vol. 51, no. 2, pp. 361–366, 1983.

[24] R. Durrett, *Probability: Theory and Examples*, 4th ed., ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 2010.

[25] Ernst, Oliver G., Mugler, Antje, Starkloff, Hans-Jörg, and Ullmann, Elisabeth, "On the convergence of generalized polynomial chaos expansions," *ESAIM: M2AN*, vol. 46, no. 2, pp. 317–339, 2012.

[26] C. C. Heyde, "On a property of the lognormal distribution," *The Journal of the Royal Statistical Society Series B (Methodological)*, vol. 25, no. 2, pp. 392–393, 1963.

[27] G. Freud, *Orthogonal Polynomials*. Pergamon Press, 1971.

[28] W. Rudin, *Real and Complex Analysis*, 3rd ed. McGraw-Hill, 1987.

[29] C. Francq and J.-M. Zakoian, *GARCH Models: Structure, Statistical Inference and Financial Applications*. Wiley, 2010. [Online]. Available: http://books.google.com/books?hl=fr{&}lr={&}id=o5g-7eaGpscC{&}pgis=1

[30] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*. Springer-Verlag, 2006. [Online]. Available: http://www.springer.com/statistics/statistical+theory+and+methods/book/978-0-387-97429-3?changeHeader

[31] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation," *Econometrica*, vol. 50, no. 4, pp. 987–1007, 1982. [Online]. Available: http://www.jstor.org/stable/1912773

[32] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, vol. 31, no. 3, pp. 307–327, 1986. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0304407686900631

[33] T. Bollerslev, R. F. Engle, and J. M. Wooldridge, "A capital asset pricing model with time varying covariances," *Journal of Political Economy*, vol. 96, pp. 116–131, 1988.

[34] R. F. Engle and F. K. Kroner, "Multivariate simultaneous generalized ARCH," *Econometric Theory*, vol. 11, pp. 122–150, 1995.

[35] T. Bollerslev, "Modelling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH model," *Review of Economics and Statistics*, vol. 72, no. 3, pp. 498–505, 1990.

[36] Y. K. Tse and A. K. C. Tsui, "A multivariate GARCH with time-varying correlations," *Journal of Business and Economic Statistics*, vol. 20, pp. 351–362, 2002.

[37] R. F. Engle, "Dynamic conditional correlation: a simple class of multivariate GARCH models," *Journal of Business and Economic Statistics*, vol. 20, pp. 339–350, 2002.

[38] F. K. Kroner and V. K. Ng, "Modelling asymmetric comovements of asset returns," *The Review of Financial Studies*, vol. 11, pp. 817–844, 1998.

[39] S. J. Taylor, "Financial returns modelled by the product of two stochastic processes, a study of daily sugar prices," in *Time series analysis: theory and practice I*, B. D. Anderson, Ed., 1982, pp. 1961–1979.

[40] A. C. Harvey, E. Ruiz, and N. Shephard, "Multivariate stochastic variance models," *Review of Economic Studies*, vol. 61, pp. 247–264, 1994.

[41] L. C. G. Rogers and D. Williams, *Diffusions, Markov Processes, and Martingales*, 2nd ed. Cambridge University Press, 2000, vol. 1.