

Discrete-time signatures and randomness in reservoir computing

Christa Cuchiero, Lukas Gonon, Lyudmila Grigoryeva, Juan-Pablo Ortega, and Josef Teichmann

Abstract—A new explanation of geometric nature of the reservoir computing phenomenon is presented. Reservoir computing is understood in the literature as the possibility of approximating input/output systems with randomly chosen recurrent neural systems and a trained linear readout layer. Light is shed on this phenomenon by constructing what is called strongly universal reservoir systems as random projections of a family of state-space systems that generate Volterra series expansions. This procedure yields a state-affine reservoir system with randomly generated coefficients in a dimension that is logarithmically reduced with respect to the original system. This reservoir system is able to approximate any element in the fading memory filters class just by training a different linear readout for each different filter. Explicit expressions for the probability distributions needed in the generation of the projected reservoir system are stated and bounds for the committed approximation error are provided.

Index Terms—Reservoir computing, recurrent neural network, state-affine system, SAS, signature state-affine system, SigSAS, echo state network, ESN, Johnson-Lindenstrauss Lemma, Volterra series, machine learning.

I. INTRODUCTION

MANY dynamical problems in engineering, signal processing, forecasting, time series analysis, recurrent neural networks, or control theory can be described using *input/output (IO) systems*. These mathematical objects establish a functional link that describes the relation between the time evolution of one or several explanatory variables (the input) and a second collection of dependent or explained variables (the output).

A generic question in all those fields is to determine the IO system underlying an observed phenomenon. This is the so called *system identification problem*. For this purpose, first principles coming from physics or chemistry can be invoked, when either these are known or the setup is simple enough to apply them. In complex situations, in which access to all the variables that determine the behavior of the systems is difficult or impossible, or when a precise mathematical relation between input and output is not known, it has proved more efficient to carry out the system identification using generic families of models with strong approximation abilities that are estimated using observed data. This approach, that we refer to as *empirical system identification*, has been developed using different techniques coming simultaneously from engineering, statistics, and computer science.

In this paper, we focus on a particularly promising strategy for empirical system identification known as *reservoir computing (RC)*. Reservoir computing capitalizes on the revolutionary idea that there are learning systems that attain universal approximation properties without the need to estimate all their parameters using, for instance, supervised learning. More specifically, RC can be seen as a recurrent neural networks approach to model IO systems *using state-space representations in which*

- the state equation is randomly generated, sometimes with sparsity features, and
- only the (usually very simple) functional form of the observation equation is tailored to the specific problem using observed data.

RC can be found in the literature under other denominations like *Liquid State Machines* [1]–[5] and is represented by various learning paradigms, with *Echo State Networks (ESNs)* [6]–[8] being a particularly important example.

RC has shown superior performance in many forecasting and classification engineering tasks (see [9]–[12] and references therein) and has shown unprecedented abilities in the learning of the attractors of complex nonlinear infinite dimensional dynamical systems [8], [13]–[15]. Additionally, RC implementations with dedicated hardware have been designed and built (see, for instance, [16]–[24]) that exhibit information processing speeds that largely outperform standard Turing-type computers.

The most far-reaching and radical innovation in the RC approach is the use of untrained, randomly generated, and sometimes sparse state maps. This circumvents well-known difficulties in the training of generic recurrent neural networks arising bifurcation phenomena [25], which, despite recent progress in the regularization and training of deep RNN structures (see, for instance [26]–[28], and references therein), render classical gradient descent methods non-convergent. Randomization has already been successfully applied in a static setup using neural networks with randomized weights, in particular in seminal works on random feature models [29] and Extreme Learning Machines [30]. In the dynamic context, an important result in [31] shows that randomly drawn ESNs can be trained by exclusively optimizing a linear readout using generic one-dimensional observations of a given invertible and differentiable dynamical system to produce dynamics that are topologically conjugate to that given system; in other words, randomly generated ESNs are capable of learning the attractors of invertible dynamical systems. More generally, the approximation capabilities of randomly generated ESNs

C. Cuchiero is affiliated with the University of Vienna, Austria. L. Gonon is with the Ludwig-Maximilians-Universität München, Germany. L. Grigoryeva is with the Universität Konstanz, Germany. J.-P. Ortega is with the Universität Sankt Gallen, Switzerland and the Centre National de la Recherche Scientifique (CNRS), France. J. Teichmann's affiliation is ETH Zürich, Switzerland.

have been established in [32] in the more general setup of IO systems. There, approximation bounds have been provided in terms of their architecture parameters.

In this paper, we provide an additional insight on the randomization question for another family of RC systems, namely, for the non-homogeneous state-affine systems (SAS). These systems have been introduced and proved to be universal approximants in [33], [34]. We here show that they also have this universality property when they are *randomly* generated. The approach pursued in this work is considerably different from the one in the above-cited references and is based on the following steps. First, we consider causal and time-invariant analytic filters with semi-infinite inputs. The Taylor series expansion of these objects coincide with what is known as their *Volterra series representation*. Second, we show that the truncated Volterra series representation (whose associated truncation error can be quantified) admits a state-space representation with linear readouts in a (potentially) high-dimensional adequately constructed tensor space. We refer to this system as the *signature state-affine system* (SigSAS): on the one hand, it belongs to the SAS family and, on the other hand, it shares two fundamental properties with the well-known signature process from the theory of rough paths [35]–[38]. First, the solutions of the SigSAS system fully characterize the input sequences and, second, any (sufficiently regular) IO system can be written as a linear map of the signature. These properties have been exploited in a continuous-time setup in [39]. Finally, we use the Johnson-Lindenstrauss Lemma [40] to prove that a random projection of the SigSAS system yields a smaller dimensional SAS system with random matrix coefficients (that can be chosen to be sparse) that approximates the original system. Moreover, this constructive procedure gives us full knowledge of the law that needs to be used to draw the entries of the low-dimensional SAS approximating system, without ever having to use the original large dimensional SigSAS, which amounts to a form of information compression with efficient reconstruction in this setup [41]. An important feature of the dimension reduced randomly drawn SAS system is that it serves as a universal approximator for any reasonably behaved IO system and that only the linear output layer that is applied to it depends on the individual system that needs to be learnt. We refer to this feature as the *strong universality* property.

This approach to the approximation problem in recurrent neural networks using randomized systems provides a new explanation of geometric nature of the reservoir computing phenomenon. The results in the following pages show that *randomly generated SAS reservoir systems approximate well any sufficiently regular IO system just by tuning a linear readout because they coincide with an error-controlled random projection of a higher dimensional Volterra series expansion of that system.*

II. TRUNCATED VOLTERRA REPRESENTATIONS OF ANALYTIC FILTERS

We start by describing the setup that we shall be working on, together with the main approximation tool which we will be using later on in the paper, namely, Volterra series expansions.

Details on the concepts introduced in the following paragraphs can be found in, for instance, [42], [43], and references therein.

All along the paper, the symbol $\mathbb{Z}$ denotes the set of all integers and $\mathbb{Z}_-$ stands for the set of negative integers with the zero element included. Let $D_d \subset \mathbb{R}^d$ and $D_m \subset \mathbb{R}^m$. We refer to the maps of the type $U : (D_d)^{\mathbb{Z}} \rightarrow (D_m)^{\mathbb{Z}}$ between infinite sequences with values in D_d and D_m , respectively, as *filters*, *operators*, or discrete-time *input/output systems*, and to those like $H : (D_d)^{\mathbb{Z}} \rightarrow D_m$ (or $H : (D_d)^{\mathbb{Z}_-} \rightarrow D_m$) as $\mathbb{R}^m$ -valued *functionals*. These definitions will be sometimes extended to accommodate situations where the domains and the targets of the filters are not necessarily product spaces but just arbitrary subsets of $(\mathbb{R}^d)^{\mathbb{Z}}$ and $(\mathbb{R}^m)^{\mathbb{Z}}$ like, for instance, $\ell^\infty(\mathbb{R}^d)$ and $\ell^\infty(\mathbb{R}^m)$.

A filter $U : (D_d)^{\mathbb{Z}} \rightarrow (D_m)^{\mathbb{Z}}$ is called *causal* when for any two elements $\mathbf{z}, \mathbf{w} \in (D_d)^{\mathbb{Z}}$ that satisfy that $\mathbf{z}_\tau = \mathbf{w}_\tau$ for any $\tau \leq t$, for a given $t \in \mathbb{Z}$, we have that $U(\mathbf{z})_t = U(\mathbf{w})_t$. Let $T_\tau : (D_d)^{\mathbb{Z}} \rightarrow (D_d)^{\mathbb{Z}}$ be the *time delay* operator defined by $T_\tau(\mathbf{z})_t := \mathbf{z}_{t-\tau}$. The filter U is called *time-invariant* (TI) when it commutes with the time delay operator, that is, $T_\tau \circ U = U \circ T_\tau$, for any $\tau \in \mathbb{Z}$ (in this expression, the two operators T_τ have to be understood as defined in the appropriate sequence spaces). There is a bijection between causal and time-invariant filters and functionals. We denote by $U_H : (D_d)^{\mathbb{Z}} \rightarrow (D_m)^{\mathbb{Z}}$ (respectively, $H_U : (D_d)^{\mathbb{Z}_-} \rightarrow D_m$) the filter (respectively, the functional) associated to the functional $H : (D_d)^{\mathbb{Z}_-} \rightarrow D_m$ (respectively, the filter $U : (D_d)^{\mathbb{Z}} \rightarrow (D_m)^{\mathbb{Z}}$). Causal and time-invariant filters are fully determined by their restriction to semi-infinite sequences, that is, $U : (D_d)^{\mathbb{Z}_-} \rightarrow (D_m)^{\mathbb{Z}_-}$, that will be denoted using the same symbol.

In most cases, we work in the situation in which D_d and D_m are compact and the sequence spaces $(D_d)^{\mathbb{Z}_-}$ and $(D_m)^{\mathbb{Z}_-}$ are endowed with the product topology. It can be shown (see [42]) that this topology is equivalent to the norm topology induced by any weighted norm defined by $\|\mathbf{z}\|_w := \sup_{t \in \mathbb{Z}_-} \{\mathbf{z}_t w_{-t}\}$, $\mathbf{z} \in (D_d)^{\mathbb{Z}_-}$, where $w : \mathbb{N} \rightarrow (0, 1]$ is an arbitrary strictly decreasing sequence (we call it *weighting sequence*) with zero limit and such that $w_0 = 1$. Filters and functionals that are continuous with respect to this topology are said to have the *fading memory property* (FMP).

A particularly important class of IO systems are those generated by *state-space systems* in which the output $\mathbf{y} \in (D_m)^{\mathbb{Z}_-}$ is obtained out of the input $\mathbf{z} \in (D_d)^{\mathbb{Z}_-}$ as the solution of the equations

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

$$(2)$$

where $F : D_N \times D_d \rightarrow D_N$ is the so-called *state map*, for some $D_N \subset \mathbb{R}^N$, $N \in \mathbb{N}$, and $h : D_N \rightarrow D_m$ is the *readout* or *observation map*. When for any input $\mathbf{z} \in (D_d)^{\mathbb{Z}_-}$ there is only one output $\mathbf{y} \in (D_m)^{\mathbb{Z}_-}$ that satisfies (1)–(2), we say that this state-space system has the *echo state property* (ESP), in which case it determines a unique filter $U_h^F : (D_d)^{\mathbb{Z}_-} \rightarrow (D_m)^{\mathbb{Z}_-}$. When the ESP holds at the level of the state equation (1), then it determines another filter $U^F : (D_d)^{\mathbb{Z}_-} \rightarrow (D_N)^{\mathbb{Z}_-}$ and then $U_h^F = h(U^F)$. The filters U_h^F and U^F , when they exist, are automatically causal

and TI (see [42]). The continuity and the differentiability properties of the state and observation maps F and h imply continuity and differentiability for U_h^F and U^F under very general hypotheses; see [43] for an in-depth study of this question.

We denote by $\|\cdot\|$ the Euclidean norm if not stated otherwise and use the symbol $\|\cdot\|$ for the operator norm with respect to the 2-norms in the target and the domain spaces. Additionally, for any $\mathbf{z} \in (\mathbb{R}^d)^{\mathbb{Z}_-}$ we define p -norms as $\|\mathbf{z}\|_p := \left(\sum_{t \in \mathbb{Z}_-} \|\mathbf{z}_t\|^p \right)^{1/p}$, for $1 \leq p < \infty$, and $\|\mathbf{z}\|_\infty := \sup_{t \in \mathbb{Z}_-} \{\|\mathbf{z}_t\|\}$ for $p = \infty$. Given $M > 0$, we denote by $K_M := \{\mathbf{z} \in (\mathbb{R}^d)^{\mathbb{Z}_-} \mid \|\mathbf{z}_t\| \leq M \text{ for all } t \in \mathbb{Z}_-\}$. It is easy to see that $K_M = \overline{B_M} \subset \ell^\infty(\mathbb{R}^d)$, with $B_M := B_{\|\cdot\|_\infty}(\mathbf{0}, M)$ and $\ell^\infty(\mathbb{R}^d) := \{\mathbf{z} \in (\mathbb{R}^d)^{\mathbb{Z}_-} \mid \|\mathbf{z}\|_\infty < \infty\}$. Additionally, let $\tilde{B}_M := B_M \cap \ell^1_-(\mathbb{R}^d)$ with $\ell^1_-(\mathbb{R}^d) := \{\mathbf{z} \in (\mathbb{R}^d)^{\mathbb{Z}_-} \mid \|\mathbf{z}\|_1 < \infty\}$. We will use the same symbol $\tilde{B}_M$ whenever $d = 1$ in the sequel. The following statement is the main approximation result that will be used in the paper.

Theorem II.1. *Let $M, L > 0$ and let $U : K_M \subset \ell^\infty(\mathbb{R}^d) \rightarrow K_L \subset \ell^\infty(\mathbb{R}^m)$ be a causal and time-invariant fading memory filter whose restriction $U|_{B_M}$ is analytic as a map between open sets in the Banach spaces $\ell^\infty(\mathbb{R}^d)$ and $\ell^\infty(\mathbb{R}^m)$ and satisfies $U(\mathbf{0}) = \mathbf{0}$. Then, for any $\mathbf{z} \in \tilde{B}_M$ there exists a Volterra series representation of U given by*

$$U(\mathbf{z})_t = \sum_{j=1}^{\infty} \sum_{m_1=-\infty}^0 \cdots \sum_{m_j=-\infty}^0 g_j(m_1, \dots, m_j)(\mathbf{z}_{m_1+t} \otimes \cdots \otimes \mathbf{z}_{m_j+t}), \quad (3)$$

with $t \in \mathbb{Z}_-$ and where the map $g_j : (\mathbb{Z}_-)^j \rightarrow L(\mathbb{R}^d \otimes \cdots \otimes \mathbb{R}^d, \mathbb{R}^m)$ is given by

$$g_j(m_1, \dots, m_j)(\mathbf{e}_{i_1} \otimes \cdots \otimes \mathbf{e}_{i_j}) = \frac{1}{j!} D^j H_U(\mathbf{0})(\mathbf{e}_{m_1}^{i_1}, \dots, \mathbf{e}_{m_j}^{i_j}), \quad (4)$$

where $\{\mathbf{e}_1, \dots, \mathbf{e}_d\}$ is the canonical basis of $\mathbb{R}^d$ and the sequences $\mathbf{e}_{m_k}^{i_l} \in \ell^\infty(\mathbb{R}^d)$ are defined by:

$$(\mathbf{e}_{m_k}^{i_l})_t := \begin{cases} \mathbf{e}_{i_l} \in \mathbb{R}^d, & \text{if } t = m_k, \\ \mathbf{0}, & \text{otherwise.} \end{cases}$$

Moreover, there exists a monotonically decreasing sequence w^U with zero limit such that, for any $p, l \in \mathbb{N}$,

$$\left\| U(\mathbf{z})_t - \sum_{j=1}^p \sum_{m_1=-l}^0 \cdots \sum_{m_j=-l}^0 g_j(m_1, \dots, m_j)(\mathbf{z}_{m_1+t} \otimes \cdots \otimes \mathbf{z}_{m_j+t}) \right\| \leq w_l^U + L \left(1 - \frac{\|\mathbf{z}\|_\infty}{M} \right)^{-1} \left(\frac{\|\mathbf{z}\|_\infty}{M} \right)^{p+1}. \quad (5)$$

A. The signature state-affine system (SigSAS)

We now show that the filter obtained out of the truncated Volterra series expansion in the expression (5) can be written down as the unique solution of a non-homogeneous state-affine system (SAS) with linear readouts that, as we shall show in Section II-B, has particularly strong universal approximation properties. We first briefly recall how the SAS family is constructed.

Let $\alpha = (\alpha_1, \dots, \alpha_d)^\top \in \mathbb{N}^d$ and $\mathbf{z} = (z_1, \dots, z_d)^\top \in \mathbb{R}^d$. We define the monomials $\mathbf{z}^\alpha := z_1^{\alpha_1} \cdots z_d^{\alpha_d}$. Let $N_1, N_2 \in \mathbb{N}$

and let $\mathbb{M}_{N_1, N_2}[\mathbf{z}]$ be the space of polynomials in $\mathbf{z} \in \mathbb{R}^d$ with matrix coefficients in $\mathbb{M}_{N_1, N_2}$, that is, the set of elements p of the form

$$p(\mathbf{z}) = \sum_{\alpha \in V_p} \mathbf{z}^\alpha A_\alpha,$$

with $V_p \subset \mathbb{N}^d$ a finite subset and $A_\alpha \in \mathbb{M}_{N_1, N_2}$ the matrix coefficients. A *state-affine system* (SAS) is given by

$$\begin{cases} \mathbf{x}_t = p(\mathbf{z}_t)\mathbf{x}_{t-1} + q(\mathbf{z}_t), \\ \mathbf{y}_t = W\mathbf{x}_t, \end{cases} \quad (6)$$

$p \in \mathbb{M}_{N, N}[\mathbf{z}]$, $q \in \mathbb{M}_{N, 1}[\mathbf{z}]$ are polynomials with matrix and vector coefficients, respectively, and $W \in \mathbb{M}_{m, N}$. If we consider inputs in the set K_M and the p and q in the state-space system (6) such that

$$\begin{aligned} M_p &:= \sup_{\mathbf{z} \in \overline{B_{\|\cdot\|}(\mathbf{0}, M)}} \{\|p(\mathbf{z})\|\} < 1, \\ M_q &:= \sup_{\mathbf{z} \in \overline{B_{\|\cdot\|}(\mathbf{0}, M)}} \{\|q(\mathbf{z})\|\} < \infty, \end{aligned}$$

where $\overline{B_{\|\cdot\|}(\mathbf{0}, M)}$ denotes the closed ball in $\mathbb{R}^d$ of radius M and center $\mathbf{0}$ with respect to the Euclidean norm, then a unique state-filter $U^{p, q} : K_M \rightarrow K_L$ can be associated to it, with $L := M_q/(1 - M_p)$. It has been shown in [33], [34] that SAS systems are universal approximants in the fading memory and in the L^p -integrable categories in the sense that given a filter in any of those two categories, there exists a SAS system of type (6) that uniformly or in the L^p -sense approximates it.

The signature state-affine system that we construct in this section exhibits what we call the *strong universality property*. This means that the state equation for this state-space representation is *the same* for any fading memory filter that is being approximated, and it is only the linear readout that changes. In other words, we provide a result that yields the approximation (as accurate as desired) of any fading memory IO system, as the linear readout of the solution of a fixed non-homogeneous SAS system that *does not depend on the filter being approximated*.

Since the important property that we just described is reminiscent of an analogous feature of the signature process in the context of the representation of the solutions of controlled stochastic differential equations [39], we shall refer to this state system as the *signature SAS* (SigSAS) system.

Before we proceed, we need to introduce some notation. First, for any $l, d \in \mathbb{N}$, we denote by $T^l(\mathbb{R}^d)$ the space of tensors of order l on $\mathbb{R}^d$, that is,

$$T^l(\mathbb{R}^d) := \left\{ \sum_{i_1, \dots, i_l=1}^d a_{i_1, \dots, i_l} \mathbf{e}_{i_1} \otimes \cdots \otimes \mathbf{e}_{i_l} \mid a_{i_1, \dots, i_l} \in \mathbb{R} \right\}.$$

The tensor space $T^l(\mathbb{R}^d)$ will be understood as a normed space with a crossnorm [44] that we shall leave unspecified for the time being. We shall be using an *order lowering map* $\pi_l : T^{l+1}(\mathbb{R}^d) \rightarrow T^l(\mathbb{R}^d)$ that, for any vector $\mathbf{v} := \sum_{i_1, \dots, i_{l+1}=1}^d a_{i_1, \dots, i_{l+1}} \mathbf{e}_{i_1} \otimes \cdots \otimes \mathbf{e}_{i_{l+1}} \in T^{l+1}(\mathbb{R}^d)$, is defined as,

$$\pi_l(\mathbf{v}) := \sum_{i_2, \dots, i_{l+1}=1}^d a_{1, i_2, \dots, i_{l+1}} \mathbf{e}_{i_2} \otimes \cdots \otimes \mathbf{e}_{i_{l+1}} \in T^l(\mathbb{R}^d).$$

The order lowering map is linear and its operator norm satisfies that $\|\pi_l\| = 1$.

We shall restrict the presentation to one-dimensional inputs, that is, we consider input sequences $\mathbf{z} \in K_M \subset \ell^\infty(\mathbb{R})$. Now, for fixed $l, p \in \mathbb{N}$, we define for any $\mathbf{z} \in K_M$ and $t \in \mathbb{Z}_-$,

$$\tilde{\mathbf{z}}_t := \sum_{i=1}^{p+1} z_t^{i-1} \mathbf{e}_i \in \mathbb{R}^{p+1} \quad \text{and} \quad \hat{\mathbf{z}}_t := \tilde{\mathbf{z}}_{t-l} \otimes \cdots \otimes \tilde{\mathbf{z}}_t. \quad (7)$$

Note that $\tilde{\mathbf{z}}_t$ is the *Vandermonde vector* [45] associated to z_t and that $\hat{\mathbf{z}}_t$ is a tensor in $T^{l+1}(\mathbb{R}^{p+1})$ whose components in the canonical basis are all the monomials on the variables $z_t, \dots, z_{t-l}$ that contain powers up to order p in each of those variables. Indeed,

$$\hat{\mathbf{z}}_t = \sum_{i_1, \dots, i_{l+1}=1}^{p+1} z_{t-l}^{i_1-1} \cdots z_t^{i_{l+1}-1} \mathbf{e}_{i_1} \otimes \cdots \otimes \mathbf{e}_{i_{l+1}} \in T^{l+1}(\mathbb{R}^{p+1}).$$

Finally, given $I_0 \subset \{1, \dots, p+1\}$ a subset of cardinality higher than 1 that contains the element 1, we define:

$$\hat{\mathbf{z}}_t^0 = \sum_{i \in I_0} z_t^{i-1} \underbrace{\mathbf{e}_1 \otimes \cdots \otimes \mathbf{e}_1}_{l\text{-times}} \otimes \mathbf{e}_i \in T^{l+1}(\mathbb{R}^{p+1}). \quad (8)$$

The next proposition introduces the SigSAS state system for fixed $l, p \in \mathbb{N}$, whose solution is used later on in Theorem II.4 to represent the truncated Volterra series expansions in Theorem II.1 of polynomial degree p and lag $-l$ (see expression (5)).

Proposition II.2 (The SigSAS system). *Let $M > 0$ and let $l, p \in \mathbb{N}$. Let $0 < \lambda < \min\{1, \frac{1-M}{1-M^{p+1}}\}$. Consider the state system with uniformly bounded scalar inputs in $K_M = [-M, M]^{\mathbb{Z}_-}$ and states in $T^{l+1}(\mathbb{R}^{p+1})$ given by the recursion*

$$\mathbf{x}_t = \lambda \pi_l(\mathbf{x}_{t-1}) \otimes \tilde{\mathbf{z}}_t + \hat{\mathbf{z}}_t^0. \quad (9)$$

This state equation is induced by the state map $F_{\lambda, l, p}^{\text{SigSAS}} : T^{l+1}(\mathbb{R}^{p+1}) \times \mathbb{R} \rightarrow T^{l+1}(\mathbb{R}^{p+1})$ defined by

$$F_{\lambda, l, p}^{\text{SigSAS}}(\mathbf{x}, z) := \lambda \pi_l(\mathbf{x}) \otimes \tilde{\mathbf{z}} + \hat{\mathbf{z}}^0, \quad (10)$$

which is a contraction in the state variable with contraction constant

$$\lambda \tilde{M} < 1, \quad \text{where} \quad \tilde{M} := \frac{1 - M^{p+1}}{1 - M} = \sum_{j=0}^p M^j, \quad (11)$$

and hence restricts to a map $F_{\lambda, l, p}^{\text{SigSAS}} : \overline{B_{\|\cdot\|}}(\mathbf{0}, L) \times [-M, M] \rightarrow \overline{B_{\|\cdot\|}}(\mathbf{0}, L)$, with

$$L := \tilde{M} / (1 - \lambda \tilde{M}). \quad (12)$$

This state system has the echo state and the fading memory properties and its continuous, time-invariant, and causal associated filter $U_{\lambda, l, p}^{\text{SigSAS}} : K_M \rightarrow K_L \subset T^{l+1}(\mathbb{R}^{p+1})$ is given by:

$$\begin{aligned} U_{\lambda, l, p}^{\text{SigSAS}}(\mathbf{z})_t &= \frac{\lambda^{l+1}}{1 - \lambda} \hat{\mathbf{z}}_t \\ &+ \lambda^l \underbrace{\pi_l(\pi_l(\cdots \pi_l(\hat{\mathbf{z}}_{t-l}^0) \otimes \tilde{\mathbf{z}}_{t-(l-1)}) \otimes \cdots) \otimes \tilde{\mathbf{z}}_{t-1}}_{l\text{-times}} \otimes \tilde{\mathbf{z}}_t + \\ &\cdots + \lambda \pi_l(\hat{\mathbf{z}}_{t-1}^0) \otimes \tilde{\mathbf{z}}_t + \hat{\mathbf{z}}_t^0. \end{aligned} \quad (13)$$

Remark II.3. The state equation (9) is indeed a SAS with states defined in $T^{l+1}(\mathbb{R}^{p+1})$ as it has the same form as the first equality in (6). Indeed, this equation can be written as $\mathbf{x}_t = p(z_t)\mathbf{x}_{t-1} + q(z_t)$ with $p(z_t)$ and $q(z_t)$ the polynomials in z_t with coefficients in $L(T^{l+1}(\mathbb{R}^{p+1}), T^{l+1}(\mathbb{R}^{p+1}))$ and $T^{l+1}(\mathbb{R}^{p+1})$, respectively, given by:

$$\begin{aligned} p(z_t)\mathbf{x}_{t-1} &:= \lambda \pi_l(\mathbf{x}_{t-1}) \otimes \tilde{\mathbf{z}}_t = \sum_{i=1}^{p+1} z_t^{i-1} (\lambda \pi_l(\mathbf{x}_{t-1}) \otimes \mathbf{e}_i), \\ q(z_t) &:= \hat{\mathbf{z}}_t^0 = \sum_{i \in I_0} z_t^{i-1} \mathbf{e}_1 \otimes \cdots \otimes \mathbf{e}_1 \otimes \mathbf{e}_i. \end{aligned}$$

B. The SigSAS approximation theorem

As we already pointed out, $\hat{\mathbf{z}}_t$ is a vector in $T^{l+1}(\mathbb{R}^{p+1})$ whose components in the canonical basis are all the monomials on the variables $z_t, \dots, z_{t-l}$ that contain powers up to order p in each of those variables. Moreover, it is easy to see that all the other summands in the expression (13) of the filter $U_{\lambda, l, p}^{\text{SigSAS}}$ are proportional (with a positive constant) to monomials already contained in $\hat{\mathbf{z}}_t$. This implies the existence of a linear map $A_{\lambda, l, p} \in L(T^{l+1}(\mathbb{R}^{p+1}), T^{l+1}(\mathbb{R}^{p+1}))$ with an invertible matrix representation with non-negative entries such that

$$U_{\lambda, l, p}^{\text{SigSAS}}(\mathbf{z})_t = A_{\lambda, l, p} \hat{\mathbf{z}}_t. \quad (14)$$

In the sequel we will denote the matrix representation of $A_{\lambda, l, p}$ using the same symbol $A_{\lambda, l, p} \in \mathbb{M}_{N, N}$, $N := (p+1)^{l+1}$. This observation, together with Theorem II.1, can be used to prove the following result.

Theorem II.4. *Let $M, L > 0$ and let $U : K_M \subset \ell^\infty(\mathbb{R}) \rightarrow K_L \subset \ell^\infty(\mathbb{R}^m)$ be a causal and time-invariant fading memory filter whose restriction $U|_{B_M}$ is analytic as a map between open sets in the Banach spaces $\ell^\infty(\mathbb{R})$ and $\ell^\infty(\mathbb{R}^m)$ and satisfies $U(\mathbf{0}) = \mathbf{0}$. Then, there exists a monotonically decreasing sequence w^U with zero limit such that, for any $p, l \in \mathbb{N}$, and any $0 < \lambda < \min\{1, \frac{1-M}{1-M^{p+1}}\}$, there exists a linear map $W \in L(T^{l+1}(\mathbb{R}^{p+1}), \mathbb{R}^m)$ such that, for any $\mathbf{z} \in \tilde{B}_M$:*

$$\|U(\mathbf{z})_t - W U_{\lambda, l, p}^{\text{SigSAS}}(\mathbf{z})_t\| \leq w_t^U + L \left(1 - \frac{\|\mathbf{z}\|_\infty}{M}\right)^{-1} \left(\frac{\|\mathbf{z}\|_\infty}{M}\right)^{p+1}. \quad (15)$$

Remark II.5. Theorem II.4 establishes the strong universality of the SigSAS system in the sense that the state equation of this system is the same for any fading memory filter U that is being approximated, and it is only the linear readout that changes. Nevertheless, we emphasize that the quality of the approximation is *not* filter independent, as the decreasing sequence w^U in the bound (15) depends on how fast the filter U “forgets” past inputs.

Remark II.6. The analyticity hypothesis in the statement of Theorem II.4 can be dropped by using the fact that finite order and finite memory Volterra series are universal approximators in the fading memory category (see [46] and [43, Theorem 31]). In that situation, the bound for the truncation error in (15) does not necessarily apply anymore, in particular its second summand, which is intrinsically linked to analyticity. A generalized bound can be formulated in that case using arguments along the lines of those found in [32].

III. JOHNSON-LINDENSTRAUSS REDUCTION OF THE SIGSAS REPRESENTATION

The price to pay for the strong universality property exhibited by the signature state-affine system that we constructed in the previous section, is the potentially large dimension of the tensor space in which this state-space representation is defined. In this section we concentrate on this problem by proposing a dimension reduction strategy which consists in using the random projections in the Johnson-Lindenstrauss Lemma [40] in order to construct a smaller dimensional SAS system with random matrix coefficients (that can be chosen to be sparse). The results contained in the next subsections quantify the increase in approximation error committed when applying this dimensionality reduction strategy.

We start by introducing the Johnson-Lindenstrauss (JL) Lemma [40] and some properties that are needed later on in the presentation. Following this, we spell out how to use it in the dimension reduction of state-space systems in general and of the SigSAS representation in particular.

A. The JL Lemma and approximate projections

Given a N -dimensional Hilbert space $(V, \langle \cdot, \cdot \rangle)$ and Q a n -point subset of V , the *Johnson-Lindenstrauss (JL) Lemma* [40] guarantees, for any $0 < \epsilon < 1$, the existence of a linear map $f : V \rightarrow \mathbb{R}^k$, with $k \in \mathbb{N}$ satisfying

$$k \geq \frac{24 \log n}{3\epsilon^2 - 2\epsilon^3}, \quad (16)$$

that respects ϵ -approximately the distances between the points in the set Q . More specifically,

$$(1-\epsilon) \|\mathbf{v}_1 - \mathbf{v}_2\|^2 \leq \|f(\mathbf{v}_1) - f(\mathbf{v}_2)\|^2 \leq (1+\epsilon) \|\mathbf{v}_1 - \mathbf{v}_2\|^2, \quad (17)$$

for any $\mathbf{v}_1, \mathbf{v}_2 \in Q$. The norm $\|\cdot\|$ in $\mathbb{R}^k$ comes from an inner product that makes it into a Hilbert space or, in other words, it satisfies the parallelogram identity. This remarkable result is even more so in connection with further developments that guarantee that the linear map f can be randomly chosen [47]–[49] and, moreover, within a family of sparse transformations [50], [51] (see also [52]).

In the developments in this paper, we use the original version of this result in which the JL map f is realized by a matrix $A \in \mathbb{M}_{k,N}$ whose entries are such that

$$A_{ij} \sim N(0, 1/k). \quad (18)$$

It can be shown that with this choice, the probability of the relation (17) to hold for any pair of points in Q is bounded below by $1/n$.

Lemma III.1. *Let $(V, \|\cdot\|)$ be a normed space and let Q be a (finite or infinite countable) subset of V . Define $\|\cdot\|_Q : \text{span}\{Q\} \rightarrow \mathbb{R}_+$ by*

$$\|\mathbf{v}\|_Q := \inf \left\{ \sum_{j=1}^{\text{Card } Q} |\lambda_j| \mid \sum_{j=1}^{\text{Card } Q} \lambda_j \mathbf{v}_j = \mathbf{v}, \mathbf{v}_j \in Q \right\}.$$

(i) $\|\cdot\|_Q$ defines a seminorm in $\text{span}\{Q\}$. If

$$M_Q := \sup \{\|\mathbf{v}_i\| \mid \mathbf{v}_i \in Q\} \quad (19)$$

is finite, then $\|\cdot\|_Q$ is a norm.

(ii) $\|\mathbf{v}\| \leq \|\mathbf{v}\|_Q M_Q$, for any $\mathbf{v} \in \text{span}\{Q\}$.

(iii) Let Q_1, Q_2 be subsets of V such that $Q_1 \subset Q_2$. Then $\|\mathbf{v}\|_{Q_2} \leq \|\mathbf{v}\|_{Q_1}$ for any $\mathbf{v} \in \text{span}\{Q_1\}$.

Remark III.2. If the hypothesis $M_Q < \infty$ is dropped in part (i) of Lemma III.1, then $\|\cdot\|_Q$ is in general not a norm as the following example shows. Take $V = \mathbb{R}$ and $\mathbf{v}_i = i$, $i \in \mathbb{N}$. It is easy to see that, in this setup,

$$\|1\|_Q = \inf \left\{ \frac{1}{i} \mid i \in \mathbb{N} \right\} = 0.$$

Proposition III.3. *Let Q be a set of points in the Hilbert space $(V, \langle \cdot, \cdot \rangle)$ with $M_Q := \sup \{\|\mathbf{v}_i\| \mid \mathbf{v}_i \in Q\} < \infty$ such that $-Q := \{-\mathbf{v} \mid \mathbf{v} \in Q\} = Q$. Let $\epsilon > 0$, let $f : V \rightarrow \mathbb{R}^k$ be a linear map that satisfies the Johnson-Lindenstrauss property (17) with respect to ϵ , and let $f^* : \mathbb{R}^k \rightarrow V$ the adjoint map with respect to a fixed inner product $\langle \cdot, \cdot \rangle$ in $\mathbb{R}^k$. Then,*

$$|\langle \mathbf{w}_1, (\mathbb{I}_V - f^* \circ f)(\mathbf{w}_2) \rangle| \leq \epsilon M_Q^2 \|\mathbf{w}_1\|_Q \|\mathbf{w}_2\|_Q, \quad (20)$$

for any $\mathbf{w}_1, \mathbf{w}_2 \in \text{span}\{Q\}$.

Corollary III.4. *In the hypotheses of the previous proposition, let*

$$C_Q := \inf_{c \in \mathbb{R}_+} \left\{ \|\mathbf{v}\|_Q \leq c \|\mathbf{v}\|, \text{ for all } \mathbf{v} \in \text{span}\{Q\} \right\}. \quad (21)$$

Then, for any $\mathbf{v} \in \text{span}\{Q\}$ such that $(f^ \circ f)(\mathbf{v}) \in \text{span}\{Q\}$, we have*

$$\|(\mathbb{I}_V - f^* \circ f)(\mathbf{v})\| \leq \epsilon M_Q^2 C_Q^2 \|\mathbf{v}\|. \quad (22)$$

This corollary is just a consequence of the inequality (20) that guarantees that

$$\begin{aligned} \|(\mathbb{I}_V - f^* \circ f)(\mathbf{v})\|^2 &\leq \epsilon M_Q^2 \|(\mathbb{I}_V - f^* \circ f)(\mathbf{v})\|_Q \|\mathbf{v}\|_Q \\ &\leq \epsilon M_Q^2 C_Q^2 \|(\mathbb{I}_V - f^* \circ f)(\mathbf{v})\| \|\mathbf{v}\|, \end{aligned} \quad (23)$$

which yields (22).

B. Johnson-Lindenstrauss projection of state-space dynamics

The next result shows how, when the dimension k of the target of the JL map f determined by (16) is chosen so that this map is generically surjective, then any contractive state-space system with states in the domain of f can be projected onto another one with states in its smaller dimensional image. This result also shows that if the original system has the ESP and the FMP, then so does the projected one. Additionally, it gives bounds that quantify the dynamical differences between the two systems.

Theorem III.5. *Let $F_\rho : \mathbb{R}^N \times D_d \rightarrow \mathbb{R}^N$ be a one-parameter family of continuous state maps, where $D_d \subset \mathbb{R}^d$ is a compact subset, $0 < \rho < 1$, and F_ρ is a ρ -contraction on the first component. Let Q be a n -point spanning subset of $\mathbb{R}^N$ satisfying $-Q = Q$. Let $f : \mathbb{R}^N \rightarrow \mathbb{R}^k$ be a JL map that satisfies (17) with $0 < \epsilon < 1$ where the dimension k has been chosen so that f is generically surjective. Then:*

(i) Let $F_\rho^f : \mathbb{R}^k \times D_d \rightarrow \mathbb{R}^k$ be the state map defined by:

$$F_\rho^f(\mathbf{x}, \mathbf{z}) := f(F_\rho(f^*(\mathbf{x}), \mathbf{z})),$$

for any $\mathbf{x} \in \mathbb{R}^k$ and $\mathbf{z} \in D_d$. If the parameter ρ is chosen so that

$$\rho < 1/\|\|f\|\|^2, \quad (24)$$

then F_ρ^f is a contraction on the first entry. The symbol $\|\cdot\|$ in (24) denotes the operator norm with respect to the 2-norms in $\mathbb{R}^N$ and $\mathbb{R}^k$.

- (ii) Let $V_k := f^*(\mathbb{R}^k) \subset \mathbb{R}^N$ and let $\mathcal{F}_\rho^f : V_k \times D_d \rightarrow V_k$ be the state map with states on the vector space V_k , defined by:

$$\mathcal{F}_\rho^f(\mathbf{x}, \mathbf{z}) := f^*(F_\rho^f((f^*)^{-1}(\mathbf{x}), \mathbf{z})) = f^* \circ f(F_\rho(\mathbf{x}, \mathbf{z})), \quad (25)$$

for any $\mathbf{x} \in V_k$ and $\mathbf{z} \in D_d$. If the contraction parameter satisfies (24) then $\mathcal{F}_\rho^f$ is also a contraction on the first entry. Moreover, the restricted linear map $f^* : \mathbb{R}^k \rightarrow V_k$ is a state-map equivariant linear isomorphism between F_ρ^f and $\mathcal{F}_\rho^f$.

- (iii) Suppose, additionally, that there exist two constants $C, C_f > 0$ such that the state spaces of the state maps F_ρ and F_ρ^f can be restricted as $F_\rho : \overline{B_{\|\cdot\|}}(\mathbf{0}, C) \times D_d \rightarrow \overline{B_{\|\cdot\|}}(\mathbf{0}, C)$ and $F_\rho^f : \overline{B_{\|\cdot\|}}(\mathbf{0}, C_f) \times D_d \rightarrow \overline{B_{\|\cdot\|}}(\mathbf{0}, C_f)$. Then, both F_ρ and F_ρ^f have the ESP and have unique FMP associated filters $U_\rho : (D_d)^{\mathbb{Z}^-} \rightarrow K_C$ and $U_\rho^f : (D_d)^{\mathbb{Z}^-} \rightarrow K_{C_f}$, respectively. The state map $\mathcal{F}_\rho^f : f^*(\overline{B_{\|\cdot\|}}(\mathbf{0}, C_f)) \times D_d \rightarrow f^*(\overline{B_{\|\cdot\|}}(\mathbf{0}, C_f))$ is isomorphic to the restricted version of F_ρ^f , also has the ESP and a FMP associated filter $\mathcal{U}_\rho^f : (D_d)^{\mathbb{Z}^-} \rightarrow (f^*(\overline{B_{\|\cdot\|}}(\mathbf{0}, C_f)))^{\mathbb{Z}^-}$. The state map $\mathcal{F}_\rho^f$ and the filter $\mathcal{U}_\rho^f$ are called the JL projected versions of F_ρ and U_ρ , respectively.

- (iv) In the hypotheses of the previous point, for any $\mathbf{z} \in (D_d)^{\mathbb{Z}^-}$ and $t \in \mathbb{Z}_-$:

$$\|U_\rho(\mathbf{z})_t - \mathcal{U}_\rho^f(\mathbf{z})_t\| \leq \epsilon^{1/2} C M_Q C_Q \frac{(1 + \|\|f\|\|^2)^{1/2}}{1 - \rho}, \quad (26)$$

where M_Q and C_Q are given by (19) and (21), respectively. Alternatively, it can also be shown that:

$$\|U_\rho(\mathbf{z})_t - \mathcal{U}_\rho^f(\mathbf{z})_t\| \leq \epsilon \frac{C M_Q^2 C_Q^2}{1 - \rho}. \quad (27)$$

- (v) Let $R > \max\{1/\|\|f\|\|^2, 1\}$ and set $\rho = 1/(R\|\|f\|\|^2)$. Then, the elements in the set Q can be chosen so that the bounds in (26) and (27) reduce to

$$\epsilon^{1/2} N^{3/4} C \left(1 + \|\|f\|\|^2\right)^{1/2} \frac{R\|\|f\|\|^2}{R\|\|f\|\|^2 - 1} \quad \text{and} \quad (28)$$

$$\epsilon N C \frac{R\|\|f\|\|^2}{R\|\|f\|\|^2 - 1}, \quad (29)$$

respectively.

C. The Johnson-Lindenstrauss reduced SigSAS system

We now use the previous theorem to spell out the Johnson-Lindenstrauss projected version of SigSAS approximations and to establish error bounds analogous to those introduced in

(28) and (29). Given that Theorem III.5 is formulated using the one and the two-norms in Euclidean spaces and Proposition II.2 defines the SigSAS system on a tensor space endowed with an unspecified cross-norm, we notice that those two frameworks can be matched by using the norms $\|\cdot\|$ and $\|\cdot\|_1$ in $T^{l+1}(\mathbb{R}^{p+1})$ given by

$$\|\mathbf{v}\|^2 := \sum_{i_1, \dots, i_{l+1}=1}^{p+1} \lambda_{i_1, \dots, i_{l+1}}^2, \quad \|\mathbf{v}\|_1^2 := \sum_{i_1, \dots, i_{l+1}=1}^{p+1} |\lambda_{i_1, \dots, i_{l+1}}|,$$

with $\mathbf{v} = \sum_{i_1, \dots, i_{l+1}=1}^{p+1} \lambda_{i_1, \dots, i_{l+1}} \mathbf{e}_{i_1} \otimes \dots \otimes \mathbf{e}_{i_{l+1}}$ and $\{\mathbf{e}_{i_1} \otimes \dots \otimes \mathbf{e}_{i_{l+1}}\}_{i_1, \dots, i_{l+1} \in \{1, \dots, p+1\}}$ the canonical basis in $T^{l+1}(\mathbb{R}^{p+1})$. It is easy to check that these two norms are cross-norms and that $\|\cdot\|$ is the norm associated to the inner product defined by the extension by bilinearity of the assignment

$$\langle \mathbf{e}_{i_1} \otimes \dots \otimes \mathbf{e}_{i_{l+1}}, \mathbf{e}_{j_1} \otimes \dots \otimes \mathbf{e}_{j_{l+1}} \rangle := \delta_{i_1 j_1} \dots \delta_{i_{l+1} j_{l+1}},$$

that makes $(T^{l+1}(\mathbb{R}^{p+1}), \langle \cdot, \cdot \rangle)$ into a Hilbert space, a feature that is needed to use the Johnson-Lindenstrauss Lemma.

Corollary III.6. Let $M > 0$ and let $(F_{\lambda, l, p}^{\text{SigSAS}}, W)$ be the SigSAS system that approximates a causal and TI filter $U : K_M \rightarrow \ell^\infty(\mathbb{R}^m)$, as introduced in Theorem II.4. Let $N := (p+1)^{l+1}$, $\tilde{M}$ as in (11), and let $0 < \epsilon < 1$. Let $f : \mathbb{R}^N \rightarrow \mathbb{R}^k$ be a JL map that satisfies (17), where the dimension k has been chosen to make f generically surjective. Then, for any $R > \max\{1/\|\|f\|\|^2, 1/(\tilde{M}\|\|f\|\|^2), 1\}$, $\lambda := 1/(R\tilde{M}\|\|f\|\|^2)$, and L as in (12), there exists a JL reduced version $\mathcal{F}_{\lambda, l, p, f}^{\text{SigSAS}} : f^*(\overline{B_{\|\cdot\|}}(\mathbf{0}, L_f)) \times [-M, M] \rightarrow f^*(\overline{B_{\|\cdot\|}}(\mathbf{0}, L_f))$ of $F_{\lambda, l, p}^{\text{SigSAS}} : \overline{B_{\|\cdot\|}}(\mathbf{0}, L) \times [-M, M] \rightarrow \overline{B_{\|\cdot\|}}(\mathbf{0}, L)$, with $L_f := \tilde{M}\|\|f\|\| / (1 - \lambda\tilde{M}\|\|f\|\|^2)$, that has the ESP and a unique FMP associated filter $\mathcal{U}_{\lambda, l, p, f}^{\text{SigSAS}} : K_M \rightarrow (f^*(\overline{B_{\|\cdot\|}}(\mathbf{0}, L_f)))^{\mathbb{Z}^-}$. Moreover, we have that

$$\begin{aligned} & \|WU_{\lambda, l, p}^{\text{SigSAS}}(\mathbf{z})_t - W\mathcal{U}_{\lambda, l, p, f}^{\text{SigSAS}}(\mathbf{z})_t\| \\ & \leq \|\|W\|\| \epsilon^{\frac{1}{2}} N^{\frac{3}{4}} (1 + \|\|f\|\|^2)^{\frac{1}{2}} \frac{\tilde{M}R^2\|\|f\|\|^4}{(R\|\|f\|\|^2 - 1)^2}, \end{aligned} \quad (30)$$

$$\begin{aligned} & \|WU_{\lambda, l, p}^{\text{SigSAS}}(\mathbf{z})_t - W\mathcal{U}_{\lambda, l, p, f}^{\text{SigSAS}}(\mathbf{z})_t\| \\ & \leq \|\|W\|\| \epsilon N \frac{\tilde{M}R^2\|\|f\|\|^4}{(R\|\|f\|\|^2 - 1)^2}, \end{aligned} \quad (31)$$

for any $\mathbf{z} \in K_M$ and $t \in \mathbb{Z}_-$, and where $W := W \circ i_k \in \mathbb{M}_{m, k}$, with $i_k : f^* \circ f(T^{l+1}(\mathbb{R}^{p+1})) \hookrightarrow T^{l+1}(\mathbb{R}^{p+1})$ the inclusion.

This result shows that causal and time-invariant filters can be approximated by JL reduced SigSAS systems. The goal in the following paragraphs consists in showing that such systems are just SAS systems with randomly drawn matrix coefficients and, additionally, in precisely spelling out the law of their entries. These facts show precisely that a large class of filters can be learnt just by randomly generating a SAS and by tuning a linear readout layer for each individual filter that needs to

be approximated. We emphasize that the JL reduced randomly generated SigSAS system is the same for the entire class of FMP filters that are being approximated and that only the linear readout depends on the individual filter that needs to be learnt, which amounts to the strong universality property that we discussed in the Introduction and in Section II-A. As in Remark II.5, we recall that the quality of the approximation using a JL reduced random SigSAS system may change from filter to filter because of the dependence on the sequence w^U in the bound (15) and the presence of the linear readout W in (30) and (31).

The next statement needs the following fact that is known in the literature as *Gordon's Theorem* (see [53, Theorem 5.32] and references therein): given a random matrix $A \in \mathbb{M}_{n,m}$ with standard Gaussian IID entries, we have that

$$\mathbb{E}[\|A\|] \leq \sqrt{n} + \sqrt{m}. \quad (32)$$

Additionally, the element $\hat{\mathbf{z}}^0 \in T^{l+1}(\mathbb{R}^{p+1})$ introduced in (8) for the construction of the SigSAS system will be chosen in a specific randomized way in this case. Indeed, this time around, we replace (8) by

$$\hat{\mathbf{z}}^0 = r \sum_{i \in I_0} z^{i-1} \mathbf{e}_1 \otimes \cdots \otimes \mathbf{e}_1 \otimes \mathbf{e}_i, \quad (33)$$

where r is a Rademacher random variable that is chosen independent from all the other random variables that will appear in the different constructions. If we take in $T^{l+1}(\mathbb{R}^{p+1})$ the canonical basis in lexicographic order, the element $\hat{\mathbf{z}}^0$ can be written as the image of a linear map as

$$\hat{\mathbf{z}}^0 = r C^{I_0} (1, z, \dots, z^p)^\top, \quad \text{with} \quad (34)$$

$$C^{I_0} := \begin{pmatrix} S^c \\ \mathbb{O}_{(p+1)((p+1)^l-1), p+1} \end{pmatrix} \in \mathbb{M}_{(p+1)^{l+1}, p+1},$$

and $S^c \in \mathbb{M}_{p+1}$ a diagonal selection matrix with the elements given by $S_{ii}^c = 1$ if $i \in I_0$, and $S_{ii}^c = 0$ otherwise.

Theorem III.7. *Let $M > 0$, let $\tilde{M}$ as in (11), $l, p, k \in \mathbb{N}$, and define $N := (p+1)^{l+1}$, $N_0 := (p+1)^l$. Consider a SigSAS state map $F_{\lambda_0, l, p}^{\text{SigSAS}} : T^{l+1}(\mathbb{R}^{p+1}) \times [-M, M] \rightarrow T^{l+1}(\mathbb{R}^{p+1})$ of the type introduced in (10) and defined by choosing the non-homogeneous term $\hat{\mathbf{z}}^0$ as in (33). Let now $f : \mathbb{R}^N \rightarrow \mathbb{R}^k$ be a JL projection randomly drawn according to (18). Let $\delta > 0$ be small enough so that*

$$\lambda_0 := \frac{\delta}{2\tilde{M}} \sqrt{\frac{k}{N_0}} < \min \left\{ \frac{1}{\tilde{M}}, \frac{1}{\tilde{M}\|f\|^2}, 1 \right\}. \quad (35)$$

Then, the JL reduced version $\mathcal{F}_{\lambda_0, l, p, f}^{\text{SigSAS}}$ of $F_{\lambda_0, l, p}^{\text{SigSAS}}$ has the ESP and the FPM with probability at least $1 - \delta$ and, in the limit $N_0 \rightarrow \infty$, it is isomorphic to the family of randomly generated SAS systems $F_{\lambda_0, l, p, f}^{\text{SigSAS}}$ with states in $\mathbb{R}^k$ and given by

$$F_{\lambda_0, l, p, f}^{\text{SigSAS}}(\mathbf{x}, z) := \sum_{i=1}^{p+1} z^{i-1} A_i \mathbf{x} + B(1, z, \dots, z^p)^\top, \quad (36)$$

where $A_1, \dots, A_{p+1} \in \mathbb{M}_k$ and $B \in \mathbb{M}_{k, p+1}$ are random matrices whose entries are drawn according to:

$$(A_1)_{j,m}, \dots, (A_{p+1})_{j,m} \sim \mathcal{N}\left(0, \frac{\delta^2}{4k\tilde{M}^2}\right), \quad (37)$$

$$B_{j,m} \sim \begin{cases} \mathcal{N}\left(0, \frac{1}{k}\right) & \text{if } m \in I_0, \\ 0 & \text{otherwise.} \end{cases} \quad (38)$$

All the entries in the matrices $A_1, \dots, A_{p+1}$ are independent random variables. The entries in the matrix B are independent from each other and they are decorrelated and asymptotically independent (in the limit as $N_0 \rightarrow \infty$) from those in $A_1, \dots, A_{p+1}$.

We conclude with a result that uses in a combined manner the SigSAS Approximation Theorem II.4 with its JL reduction in Corollary III.6, as well as its SAS characterization with random coefficients in Theorem III.7. This statement shows that in order to approximate a large class of sufficiently regular FMP filters with uniformly bounded inputs, it suffices to randomly generate a common SAS system for all of them and to tune a linear readout for each different filter in that class that needs to be approximated.

Theorem III.8. *Let $M, L > 0$ and let $U : K_M \subset \ell^\infty(\mathbb{R}) \rightarrow K_L \subset \ell^\infty(\mathbb{R}^m)$ be a causal and time-invariant fading memory filter that satisfies the hypotheses in Theorem II.4. Fix now $l, p, k \in \mathbb{N}$ and $\delta > 0$ small enough so that (35) holds. Construct now the SAS system with states in $\mathbb{R}^k$ given by*

$$F_{\lambda_0, l, p, f}^{\text{SigSAS}}(\mathbf{x}, z) = \sum_{i=1}^{p+1} z^{i-1} A_i \mathbf{x} + B(1, z, \dots, z^p)^\top, \quad (39)$$

with matrix coefficients randomly generated according to the laws spelled out in (37) and (38).

If p and l are large enough, then the SAS system $F_{\lambda_0, l, p, f}^{\text{SigSAS}}$ has the ESP and the FPM with probability at least $1 - \delta$. In that case $F_{\lambda_0, l, p, f}^{\text{SigSAS}}$ has a filter $U_{\lambda_0, l, p, f}^{\text{SigSAS}}$ associated and there exists a monotonically decreasing sequence w^U with zero limit and a linear map $\bar{W} \in L(\mathbb{R}^k, \mathbb{R}^m)$ such that for any $\mathbf{z} \in \tilde{B}_M$ it holds that

$$\|U(\mathbf{z})_t - \bar{W} U_{\lambda_0, l, p, f}^{\text{SigSAS}}(\mathbf{z})_t\| \leq w_t^U + L \left(1 - \frac{\|\mathbf{z}\|_\infty}{M}\right)^{-1} \left(\frac{\|\mathbf{z}\|_\infty}{M}\right)^{p+1} + I_{l,p}, \quad (40)$$

where $I_{l,p}$ is either

$$I_{l,p} := \|W\| \epsilon^{\frac{1}{2}} N^{\frac{3}{4}} \tilde{M} \frac{\left(1 + \|f\|^2\right)^{\frac{1}{2}}}{\left(1 - \frac{\delta}{2} \sqrt{\frac{k}{N_0}}\right)^2} \quad \text{or} \quad (41)$$

$$I_{l,p} := \|W\| \epsilon N \tilde{M} \frac{1}{\left(1 - \frac{\delta}{2} \sqrt{\frac{k}{N_0}}\right)^2}.$$

In these expressions $W \in L(T^{l+1}(\mathbb{R}^{p+1}), \mathbb{R}^m)$ is a linear map such that $\bar{W} = W \circ f^$, $N = (p+1)^{l+1}$, $\tilde{M}$ is defined in (11), and $0 < \epsilon < 1$ satisfies (16) with n replaced by N .*

IV. CONCLUSION

Reservoir computing capitalizes on the remarkable fact that there are learning systems that attain universal approximation properties without requiring that all their parameters are estimated using a supervised learning procedure. These untrained parameters are most of the time randomly generated and it

is only an output layer that needs to be estimated using a simple functional prescription. This phenomenon has been explained for static (extreme learning machines [30]) and dynamic (echo state networks [31], [32]) neural paradigms and its performance has been quantified using mostly probabilistic methods.

In this paper, we have concentrated on a different class of reservoir computing systems, namely the state-affine (SAS) family. The SAS class was introduced and proved universal in [33] and we have shown here that the possibility of randomly constructing these systems and at the same time preserving their approximation properties is of geometric nature. The rationale behind our description relies on the following points:

- Any analytic filter can be represented as a Volterra series expansion. When this filter is additionally of fading memory type, the truncation error can be easily quantified.
- Truncated Volterra series admit a natural state-space representation with linear observation equation in a conveniently chosen tensor space. The state equation of this representation has a strong universality property whose unique solution can be used to approximate any analytic fading memory filter just by modifying the linear observation equation. We refer to this strongly universal filter as the SigSAS system.
- The random projections of the SigSAS system yield SAS systems with randomly generated coefficients in a potentially much smaller dimension which approximately preserve the good properties of the original SigSAS system. The loss in performance that one incurs because of the projection mechanism can be quantified using the Johnson-Lindenstrauss Lemma.

These observations collectively show that *SAS reservoir systems with randomly chosen coefficients exhibit excellent empirical performances in the learning of fading memory input/output systems because they approximately correspond to very high-degree Volterra series expansions of those systems.* That degree is actually of the order of the exponential of the dimension of the randomly generated SAS system used.

APPENDIX

A. Proof of Theorem II.1

The representation (3) is a straightforward multivariate generalization of Theorem 29 in [43]. For any $\mathbf{z} \in \tilde{B}_M$ and any $p, l \in \mathbb{N}$ define

$$U^{l,p}(\mathbf{z})_t := \sum_{j=1}^p \sum_{m_1=-l}^0 \cdots \sum_{m_j=-l}^0 g_j(m_1, \dots, m_j)(\mathbf{z}_{m_1+t} \otimes \cdots \otimes \mathbf{z}_{m_j+t}).$$

Now, for any $\mathbf{z} \in \tilde{B}_M$ and $t_1, t_2 \in \mathbb{Z}_-$ such that $t_2 \leq t_1$, define the sequence $\mathbf{z}_{t_2}^{t_1} \in \tilde{B}_M$ by $\mathbf{z}_{t_2}^{t_1} := (\dots, \mathbf{0}, \mathbf{z}_{t_2}, \dots, \mathbf{z}_{t_1})$. Additionally, for any $\mathbf{u} \in (\mathbb{R}^d)^{\mathbb{Z}_-}$ and any $\mathbf{z} \in (\mathbb{R}^d)^{\mathbb{N}^+}$, the symbol $\mathbf{u}\mathbf{z}_t^1 \in (\mathbb{R}^d)^{\mathbb{Z}_-}$, $t \in \mathbb{N}^+$, denotes the concatenation of the left-shifted vector $\mathbf{u}$ with the truncated vector $\mathbf{z}_t^1 := (\mathbf{z}_1, \dots, \mathbf{z}_t)$ obtained out of $\mathbf{z}$. With this notation, we now show (5). By the triangle inequality and the time-invariance of U , for any $\mathbf{z} \in \tilde{B}_M$ we have

$$\begin{aligned} \|U(\mathbf{z})_t - U^{l,p}(\mathbf{z})_t\| &\leq \|U(\mathbf{z})_t - U^{l,\infty}(\mathbf{z})_t\| + \|U^{l,\infty}(\mathbf{z})_t - U^{l,p}(\mathbf{z})_t\| \\ &= \left\| \sum_{j=1}^{\infty} \sum_{m_1=-\infty}^{-l-1} \cdots \sum_{m_j=-\infty}^{-l-1} g_j(m_1, \dots, m_j)(\mathbf{z}_{m_1+t} \otimes \cdots \otimes \mathbf{z}_{m_j+t}) \right\| \\ &\quad + \|U(\mathbf{z}_{-l+t}^t)_0 - U^{\infty,p}(\mathbf{z}_{-l+t}^t)_0\| \\ &= \|U(\mathbf{z}_{-\infty}^{-l-1+t} \mathbf{0}_{l+1})_0\| + \|U(\mathbf{z}_{-l+t}^t)_0 - U^{\infty,p}(\mathbf{z}_{-l+t}^t)_0\|, \end{aligned} \quad (42)$$

where the symbol $\mathbf{0}_{l+1}$ stands for a $l+1$ -tuple of the element $\mathbf{0} \in \mathbb{R}^d$. The second summand of this expression can be bounded using the Taylor bound provided in [43, Theorem 29]. As to the first summand, we shall use the input forgetting property that the filter U exhibits since, by hypothesis, has the FMP. More specifically, if we apply Theorem 6 in [43] to the FMP filter $U : K_M \rightarrow \ell_{\infty}^m(\mathbb{R}^m)$, we can conclude the existence of a monotonically decreasing sequence w^U with zero limit such that

$$\|U(\mathbf{z}_{-\infty}^{-l-1+t} \mathbf{0}_{l+1})_0\| = \|U(\mathbf{z}_{-\infty}^{-l-1+t} \mathbf{0}_{l+1})_0 - U(\mathbf{0})_0\| \leq w_l^U,$$

for any $l \in \mathbb{N}$. These two arguments substituted in (42) yield the bound in (5). ■

B. Proof of Proposition II.2

The map $F_{\lambda,l,p}^{\text{SigSAS}} : T^{l+1}(\mathbb{R}^{p+1}) \times [-M, M] \rightarrow T^{l+1}(\mathbb{R}^{p+1})$ is clearly continuous and, additionally, it is a contraction in its first component. Indeed, let $\mathbf{x}_1, \mathbf{x}_2 \in T^{l+1}(\mathbb{R}^{p+1})$ and let $z \in [-M, M]$ be arbitrary. Notice first that

$$\|\tilde{\mathbf{z}}\| = \left\| \sum_{i=1}^{p+1} z^{i-1} \mathbf{e}_i \right\| \leq 1 + M + \cdots + M^p = \frac{1 - M^{p+1}}{1 - M} =: \tilde{M}. \quad (43)$$

Now, since we are using a crossnorm in $T^{l+1}(\mathbb{R}^{p+1})$, we have that,

$$\begin{aligned} \|F_{\lambda,l,p}^{\text{SigSAS}}(\mathbf{x}_1, z) - F_{\lambda,l,p}^{\text{SigSAS}}(\mathbf{x}_2, z)\| &= \lambda \|\pi_l(\mathbf{x}_1 - \mathbf{x}_2) \otimes \tilde{\mathbf{z}}\| \\ &= \lambda \|\pi_l(\mathbf{x}_1 - \mathbf{x}_2)\| \|\tilde{\mathbf{z}}\|. \end{aligned}$$

If we use in this equality the relation (43) and the fact that $\|\pi_l\| = 1$, we can conclude that:

$$\|F_{\lambda,l,p}^{\text{SigSAS}}(\mathbf{x}_1, z) - F_{\lambda,l,p}^{\text{SigSAS}}(\mathbf{x}_2, z)\| \leq \lambda \tilde{M} \|\mathbf{x}_1 - \mathbf{x}_2\|. \quad (44)$$

The hypothesis $\lambda < (1 - M)/(1 - M^{p+1}) = 1/\tilde{M}$ implies that $F_{\lambda,l,p}^{\text{SigSAS}}$ is a contraction and establishes (11). Additionally, $\|\tilde{\mathbf{z}}\| = \left\| \sum_{i \in I_0} z^{i-1} \mathbf{e}_1 \otimes \cdots \otimes \mathbf{e}_1 \otimes \mathbf{e}_i \right\| \leq 1 + M + \cdots + M^p = \tilde{M}$, which implies that

$$\|F_{\lambda,l,p}^{\text{SigSAS}}(\mathbf{0}, z)\| \leq \tilde{M}, \quad \text{for all } z \in [-M, M], \quad (45)$$

and hence by [54, Remark 2] we can conclude that $F_{\lambda,l,p}^{\text{SigSAS}}$ restricts to a map $F_{\lambda,l,p}^{\text{SigSAS}} : \overline{B_{\|\cdot\|}}(\mathbf{0}, L) \times [-M, M] \rightarrow \overline{B_{\|\cdot\|}}(\mathbf{0}, L)$, for any $L \geq \tilde{M}/(1 - \lambda\tilde{M})$. Finally, the contractivity condition established in (44) and [43, Theorem 12], imply that the corresponding state system has the ESP and the FMP. We now show that its unique solution is given by (13). First, it is easy to see that by iterating the recursion (9) twice and three times, one obtains:

$$\begin{aligned} \mathbf{x}_t &= \lambda^2 \pi_l(\pi_l(\mathbf{x}_{t-2}) \otimes \tilde{\mathbf{z}}_{t-1}) \otimes \tilde{\mathbf{z}}_t + \lambda \pi_l(\tilde{\mathbf{z}}_{t-1}^0) \otimes \tilde{\mathbf{z}}_t + \tilde{\mathbf{z}}_t^0 \\ &= \lambda^3 \pi_l(\pi_l(\pi_l(\mathbf{x}_{t-3}) \otimes \tilde{\mathbf{z}}_{t-2}) \otimes \tilde{\mathbf{z}}_{t-1}) \otimes \tilde{\mathbf{z}}_t \\ &\quad + \lambda^2 \pi_l(\pi_l(\tilde{\mathbf{z}}_{t-2}^0) \otimes \tilde{\mathbf{z}}_{t-1}) \otimes \tilde{\mathbf{z}}_t + \lambda \pi_l(\tilde{\mathbf{z}}_{t-1}^0) \otimes \tilde{\mathbf{z}}_t + \tilde{\mathbf{z}}_t^0. \end{aligned}$$

More generally, after $l+1$ iterations one obtains,

$$\begin{aligned} \mathbf{x}_t &= \lambda^{l+1} \underbrace{\pi_l(\pi_l(\cdots (\pi_l(\mathbf{x}_{t-(l+1)}) \otimes \tilde{\mathbf{z}}_{t-l}) \otimes \cdots) \otimes \tilde{\mathbf{z}}_{t-1}) \otimes \tilde{\mathbf{z}}_t}_{l+1\text{-times}} \\ &\quad + \lambda^l \underbrace{\pi_l(\pi_l(\cdots (\pi_l(\tilde{\mathbf{z}}_{t-l}^0) \otimes \tilde{\mathbf{z}}_{t-(l-1)}) \otimes \cdots) \otimes \tilde{\mathbf{z}}_{t-1}) \otimes \tilde{\mathbf{z}}_t}_{l\text{-times}} \\ &\quad + \cdots + \lambda \pi_l(\tilde{\mathbf{z}}_{t-1}^0) \otimes \tilde{\mathbf{z}}_t + \tilde{\mathbf{z}}_t^0. \end{aligned}$$

Consequently, in order to establish (13) it suffices to show that

$$\lambda^{l+1} \underbrace{\pi_l(\pi_l(\cdots (\pi_l(\mathbf{x}_{t-(l+1)}) \otimes \tilde{\mathbf{z}}_{t-l}) \otimes \cdots) \otimes \tilde{\mathbf{z}}_{t-1}) \otimes \tilde{\mathbf{z}}_t}_{l+1\text{-times}} = \frac{\lambda^{l+1}}{1 - \lambda} \tilde{\mathbf{z}}_t. \quad (46)$$

We show this equality by writing

$$\mathbf{x}_t = \sum_{i_1, \dots, i_{l+1}=1}^{p+1} a_{i_1, \dots, i_{l+1}}^t \mathbf{e}_{i_1} \otimes \cdots \otimes \mathbf{e}_{i_{l+1}}, \quad (47)$$

for some coefficients $a_{i_1, \dots, i_{l+1}}^t \in \mathbb{R}$ that by (9) and the assumption that $1 \in I_0$, satisfy that

$$a_{1, \dots, 1}^t = \lambda a_{1, \dots, 1}^{t-1} + 1, \quad \text{for any } t \in \mathbb{Z}_-.$$

This recursion can be rewritten for any $r \in \mathbb{N}$ and $t \in \mathbb{Z}$ as

$$a_{1, \dots, 1}^t = \sum_{j=0}^{r-1} \lambda^j + \lambda^r a_{1, \dots, 1}^{t-r}.$$

Since by hypothesis the parameter $\lambda < 1$ and, additionally, we just showed that $\|\mathbf{x}_t\| \leq L$, for all $t \in \mathbb{Z}$, with $L \geq \bar{M}/(1 - \lambda\bar{M})$, this equation has a unique solution given by

$$a_{1, \dots, 1}^t = \frac{1}{1 - \lambda}, \quad \text{for all } t \in \mathbb{Z}_-. \quad (48)$$

Now, notice that using (47), we can write:

$$\begin{aligned} & \pi_l(\mathbf{x}_{t-(l+1)}) \otimes \tilde{\mathbf{z}}_{t-l} \\ &= \sum_{i_2, \dots, i_{l+1}, j_1=1}^{p+1} a_{1, \dots, 1}^{t-(l+1)} z_{t-l}^{j_1-1} \mathbf{e}_{i_2} \otimes \dots \otimes \mathbf{e}_{i_{l+1}} \otimes \mathbf{e}_{j_1}. \end{aligned}$$

If we repeat this procedure $l+1$ times, we obtain that

$$\begin{aligned} & \underbrace{\pi_l(\pi_l(\dots(\pi_l(\mathbf{x}_{t-(l+1)}) \otimes \tilde{\mathbf{z}}_{t-l}) \otimes \dots) \otimes \tilde{\mathbf{z}}_{t-1}) \otimes \tilde{\mathbf{z}}_t}_{l+1\text{-times}} \\ &= \sum_{j_1, \dots, j_{l+1}=1}^{p+1} a_{1, \dots, 1}^{t-(l+1)} z_{t-l}^{j_1-1} z_{t-(l-1)}^{j_2-1} \dots z_t^{j_{l+1}-1} \mathbf{e}_{j_1} \otimes \dots \otimes \mathbf{e}_{j_{l+1}} \\ &= a_{1, \dots, 1}^{t-(l+1)} \hat{\mathbf{z}}_t = \frac{1}{1 - \lambda} \hat{\mathbf{z}}_t, \end{aligned}$$

where the last equality is a consequence of (48). This identity proves (46). ■

C. Proof of Theorem II.4

It is a straightforward corollary of Theorem II.1 and of the expression (13) of the filter $U_{\lambda, l, p}^{\text{SigSAS}}$. The linear map W is constructed by matching the coefficients $g_j(m_1, \dots, m_j)$ of the truncated Volterra series representation of U up to polynomial degree p with the terms of the filter $U_{\lambda, l, p}^{\text{SigSAS}}(\mathbf{z})_t$ in the canonical basis of $T^{l+1}(\mathbb{R}^{p+1})$. More specifically, $W \in L(T^{l+1}(\mathbb{R}^{p+1}), \mathbb{R}^m)$ is the linear map that satisfies:

$$WU_{\lambda, l, p}^{\text{SigSAS}}(\mathbf{z})_t = \sum_{j=1}^p \sum_{m_1=-l}^0 \dots \sum_{m_j=-l}^0 g_j(m_1, \dots, m_j) z_{m_1+t} \dots z_{m_j+t}, \quad (49)$$

for any $\mathbf{z} \in K_M$, $t \in \mathbb{Z}_-$, where the right hand side of this equality is the truncated Volterra series expansion of U , available by Theorem II.1. The equality (49) does determine W because by (14), it is equivalent to:

$$\begin{aligned} & \sum_{i_1, \dots, i_{l+1}=1}^{p+1} WA_{\lambda, l, p}(\mathbf{e}_{i_1} \otimes \dots \otimes \mathbf{e}_{i_{l+1}}) z_{t-l}^{i_1-1} \dots z_t^{i_{l+1}-1} \\ &= \sum_{j=1}^p \sum_{m_1=-l}^0 \dots \sum_{m_j=-l}^0 g_j(m_1, \dots, m_j) z_{m_1+t} \dots z_{m_j+t}. \end{aligned}$$

Since this equality between polynomials has to hold for any $\mathbf{z} \in K_M$ and $t \in \mathbb{Z}_-$, we can conclude that the matrix coefficients on both sides have to coincide. This implies that, in particular, for any $i_1, \dots, i_{l+1} \in \{1, \dots, p+1\}$,

$$WA_{\lambda, l, p}(\mathbf{e}_{i_1} \otimes \dots \otimes \mathbf{e}_{i_{l+1}}) = \sum_{I_{i_1, \dots, i_{l+1}}} g_j(m_1, \dots, m_j) \in \mathbb{R}^m, \quad (50)$$

where $I_{i_1, \dots, i_{l+1}} = \{(j, m_1, \dots, m_j)\}$ is the set of indices with $j \in \{1, \dots, p\}$, $m_i \in \{-l, \dots, 0\}$, and $z_{m_1} \dots z_{m_j} = z_{-l}^{i_1-1} \dots z_0^{i_{l+1}-1}$. As (50) specifies the image of a basis by the map $WA_{\lambda, l, p}$ and $A_{\lambda, l, p}$ is invertible, then (50) and consequently (49) fully determine W . The bound in (15) is then a consequence of (49) and (5) in Theorem II.1. ■

D. Proof of Lemma III.1

(i) It is obvious that if $\mathbf{v} = \mathbf{0}$ then $\|\mathbf{v}\|_Q = 0$ and that $\|\lambda\mathbf{v}\|_Q = |\lambda| \|\mathbf{v}\|_Q$, for all $\lambda \in \mathbb{R}$ and $\mathbf{v} \in \text{span}\{Q\}$. Let now $\mathbf{w}_1, \mathbf{w}_2 \in \text{span}\{Q\}$ and $C_Q := \text{Card } Q$. Given that,

$$\begin{aligned} & \inf \left\{ \sum_{j=1}^{C_Q} |\lambda_j^1 + \lambda_j^2| \mid \sum_{j=1}^{C_Q} \lambda_j^1 \mathbf{v}_j = \mathbf{w}_1, \sum_{j=1}^{C_Q} \lambda_j^2 \mathbf{v}_j = \mathbf{w}_2, \mathbf{v}_j \in Q \right\} \\ & \geq \inf \left\{ \sum_{j=1}^{C_Q} |\lambda_j| \mid \sum_{j=1}^{C_Q} \lambda_j \mathbf{v}_j = \mathbf{w}_1 + \mathbf{w}_2, \mathbf{v}_j \in Q \right\}, \end{aligned}$$

we can conclude that

$$\begin{aligned} & \|\mathbf{w}_1 + \mathbf{w}_2\|_Q \leq \\ & \inf \left\{ \sum_{j=1}^{C_Q} |\lambda_j^1 + \lambda_j^2| \mid \sum_{j=1}^{C_Q} \lambda_j^1 \mathbf{v}_j = \mathbf{w}_1, \sum_{j=1}^{C_Q} \lambda_j^2 \mathbf{v}_j = \mathbf{w}_2, \mathbf{v}_j \in Q \right\} \\ & \leq \inf \left\{ \sum_{j=1}^{C_Q} |\lambda_j^1| + |\lambda_j^2| \mid \sum_{j=1}^{C_Q} \lambda_j^1 \mathbf{v}_j = \mathbf{w}_1, \sum_{j=1}^{C_Q} \lambda_j^2 \mathbf{v}_j = \mathbf{w}_2, \mathbf{v}_j \in Q \right\} \\ & = \|\mathbf{w}_1\|_Q + \|\mathbf{w}_2\|_Q, \end{aligned}$$

which establishes the triangle inequality and hence shows that $\|\cdot\|_Q$ is a seminorm. Suppose now that $M_Q < \infty$ and let $\mathbf{v} \in \text{span}\{Q\}$ such that $\|\mathbf{v}\|_Q = 0$. By the approximation property of the infimum, for any $\epsilon > 0$ there exist $\lambda_1, \dots, \lambda_{C_Q} \in \mathbb{R}$ such that $\sum_{j=1}^{C_Q} \lambda_j \mathbf{v}_j = \mathbf{v}$ and $0 \leq \sum_{j=1}^{C_Q} |\lambda_j| < \epsilon$. This inequality implies that

$$\|\mathbf{v}\| = \left\| \sum_{j=1}^{C_Q} \lambda_j \mathbf{v}_j \right\| \leq M_Q \sum_{j=1}^{C_Q} |\lambda_j| < M_Q \epsilon. \quad (51)$$

Since M_Q is finite and $\epsilon > 0$ can be made arbitrarily small, this inequality implies that $\|\mathbf{v}\| = 0$ and hence $\mathbf{v} = \mathbf{0}$, necessarily, which proves that $\|\cdot\|_Q$ is a norm in this case.

Since the first inequality (51) holds for any $\mathbf{v} \in \text{span}\{Q\}$, the statement in part (ii) follows (when M_Q is not finite we use the convention that $\infty \cdot 0 = \infty$). Part (iii) is obvious. ■

E. Proof of Proposition III.3

Since V and $\mathbb{R}^k$ are Hilbert spaces, the parallelogram law holds for the associated norms and hence, for any $\mathbf{v}_1, \mathbf{v}_2 \in Q$,

$$\begin{aligned} & \langle \mathbf{v}_1, \mathbf{v}_2 - f^* \circ f(\mathbf{v}_2) \rangle = \langle \mathbf{v}_1, \mathbf{v}_2 \rangle - \langle f(\mathbf{v}_1), f(\mathbf{v}_2) \rangle \\ &= \frac{1}{4} \left(\|\mathbf{v}_1 + \mathbf{v}_2\|^2 - \|\mathbf{v}_1 - \mathbf{v}_2\|^2 \right) \\ & \quad - \frac{1}{4} \left(\|f(\mathbf{v}_1) + f(\mathbf{v}_2)\|^2 - \|f(\mathbf{v}_1) - f(\mathbf{v}_2)\|^2 \right) \\ &= \frac{1}{4} \left(\|\mathbf{v}_1 - (-\mathbf{v}_2)\|^2 - \|\mathbf{v}_1 - \mathbf{v}_2\|^2 \right) \\ & \quad - \frac{1}{4} \left(\|f(\mathbf{v}_1) - f(-\mathbf{v}_2)\|^2 - \|f(\mathbf{v}_1) - f(\mathbf{v}_2)\|^2 \right) \\ & \leq \frac{\epsilon}{4} \left(\|\mathbf{v}_1 - \mathbf{v}_2\|^2 + \|\mathbf{v}_1 + \mathbf{v}_2\|^2 \right) = \frac{\epsilon}{2} \left(\|\mathbf{v}_1\|^2 + \|\mathbf{v}_2\|^2 \right), \quad (52) \end{aligned}$$

where in the inequality in the last line we used the JL property (17) together with the hypothesis $-Q = Q$. Let now $\mathbf{w}_1 = \sum_{i=1}^{\text{Card } Q} \lambda_i^1 \mathbf{v}_i$, $\mathbf{w}_2 = \sum_{i=1}^{\text{Card } Q} \lambda_i^2 \mathbf{v}_i \in \text{span}\{Q\}$. Then, by (52):

$$\begin{aligned} & |\langle \mathbf{w}_1, \mathbf{w}_2 - f^* \circ f(\mathbf{w}_2) \rangle| = \left| \sum_{i,j=1}^{\text{Card } Q} \lambda_i^1 \lambda_j^2 \langle \mathbf{v}_i, \mathbf{v}_j - f^* \circ f(\mathbf{v}_j) \rangle \right| \\ & \leq \sum_{i,j=1}^{\text{Card } Q} |\lambda_i^1| |\lambda_j^2| \frac{\epsilon}{2} \left(\|\mathbf{v}_i\|^2 + \|\mathbf{v}_j\|^2 \right) \leq \sum_{i=1}^{\text{Card } Q} |\lambda_i^1| \sum_{j=1}^{\text{Card } Q} |\lambda_j^2| \epsilon M_Q^2. \end{aligned}$$

Since this inequality holds true for any linear decomposition of $\mathbf{w}_1, \mathbf{w}_2 \in \text{span}\{Q\}$, we can take infima on its right hand side with respect to those decompositions, which clearly implies (20). ■

F. Proof of Theorem III.5

(i) We show that when condition (24) holds, then F_ρ^f is a contraction on the first entry. Let $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^k$ and let $\mathbf{z} \in D_d$, then

$$\begin{aligned} & \left\| F_\rho^f(\mathbf{x}_1, \mathbf{z}) - F_\rho^f(\mathbf{x}_2, \mathbf{z}) \right\| \\ &= \left\| f(F_\rho(f^*(\mathbf{x}_1), \mathbf{z})) - f(F_\rho(f^*(\mathbf{x}_2), \mathbf{z})) \right\| \leq \rho \|f\| \|f^*\| \|\mathbf{x}_1 - \mathbf{x}_2\|. \end{aligned}$$

The claim follows from this inequality, the equality $\|f\| = \|f^*\|$, and condition (24).

(ii) The proof is straightforward. The only point that needs to be emphasized is that $(f^*)^{-1} : V_k \rightarrow \mathbb{R}^k$ is well-defined because since f is surjective, then $f^* : \mathbb{R}^k \rightarrow V_k$ is necessarily injective.

(iii) First of all, the existence of the restricted versions of F_ρ and F_ρ^f to compact state-spaces and the fact that these maps are contractions on the first entry with contraction rates ρ and $\rho\|f\|^2$, respectively, implies by [43, Theorem 7, part (i)] that they have the ESP and associated FMP filters U_ρ and U_ρ^f . The statement about the JL-projected state map $\mathcal{F}_\rho^f$ and its associated filter $\mathcal{U}_\rho^f$ is a straightforward consequence of the fact that the restricted linear map $f^* : \mathbb{R}^k \rightarrow V_k$ is a state-map equivariant linear isomorphism between F_ρ^f and $\mathcal{F}_\rho^f$ and of the properties of this kind of maps (see, for instance, [55, Proposition 2.3]).

(iv) Let $\mathbf{z} \in (D_d)^{\mathbb{Z}_-}$ and $t \in \mathbb{Z}_-$ arbitrary. Then, using (25), we have

$$\begin{aligned} & \left\| U_\rho(\mathbf{z})_t - \mathcal{U}_\rho^f(\mathbf{z})_t \right\| = \left\| F_\rho(U_\rho(\mathbf{z})_{t-1}, \mathbf{z}_t) - \mathcal{F}_\rho(\mathcal{U}_\rho^f(\mathbf{z})_{t-1}, \mathbf{z}_t) \right\| \\ &= \left\| F_\rho(U_\rho(\mathbf{z})_{t-1}, \mathbf{z}_t) - F_\rho(\mathcal{U}_\rho^f(\mathbf{z})_{t-1}, \mathbf{z}_t) \right\| \\ &+ \left\| F_\rho(\mathcal{U}_\rho^f(\mathbf{z})_{t-1}, \mathbf{z}_t) - \mathcal{F}_\rho(\mathcal{U}_\rho^f(\mathbf{z})_{t-1}, \mathbf{z}_t) \right\| \\ &\leq \rho \left\| U_\rho(\mathbf{z})_{t-1} - \mathcal{U}_\rho^f(\mathbf{z})_{t-1} \right\| + \left\| (\mathbb{I}_N - f^* \circ f)(F_\rho(\mathcal{U}_\rho^f(\mathbf{z})_{t-1}, \mathbf{z}_t)) \right\|. \end{aligned} \quad (53)$$

The bounds in (26) and (27) are obtained by bounding the last expression in (53) in two different fashions. First, if we use (20) and the hypothesis on Q being a spanning set of $\mathbb{R}^N$, we have that:

$$\begin{aligned} & \rho \left\| U_\rho(\mathbf{z})_{t-1} - \mathcal{U}_\rho^f(\mathbf{z})_{t-1} \right\| + \left\| (\mathbb{I}_N - f^* \circ f)F_\rho(\mathcal{U}_\rho^f(\mathbf{z})_{t-1}, \mathbf{z}_t) \right\| \\ &\leq \rho \left\| U_\rho(\mathbf{z})_{t-1} - \mathcal{U}_\rho^f(\mathbf{z})_{t-1} \right\| \\ &+ \epsilon^{1/2} M_Q \left\| (\mathbb{I}_N - f^* \circ f)(F_\rho(\mathcal{U}_\rho^f(\mathbf{z})_{t-1}, \mathbf{z}_t)) \right\|_Q^{1/2} \left\| F_\rho(\mathcal{U}_\rho^f(\mathbf{z})_{t-1}, \mathbf{z}_t) \right\|_Q^{1/2} \\ &\leq \rho \left\| U_\rho(\mathbf{z})_{t-1} - \mathcal{U}_\rho^f(\mathbf{z})_{t-1} \right\| + \epsilon^{1/2} C M_Q C_Q \left(1 + \|f\|^2 \right)^{1/2}. \end{aligned} \quad (54)$$

If we now iterate the procedure in (53) on the first summand of this expression, we obtain

$$\begin{aligned} & \left\| U_\rho(\mathbf{z})_t - \mathcal{U}_\rho^f(\mathbf{z})_t \right\| = \left\| F_\rho(U_\rho(\mathbf{z})_{t-1}, \mathbf{z}_t) - \mathcal{F}_\rho(\mathcal{U}_\rho^f(\mathbf{z})_{t-1}, \mathbf{z}_t) \right\| \\ &\leq \rho \left(\left\| U_\rho(\mathbf{z})_{t-2} - \mathcal{U}_\rho^f(\mathbf{z})_{t-2} \right\| + \epsilon^{1/2} C M_Q C_Q \left(1 + \|f\|^2 \right)^{1/2} \right) \\ &+ \epsilon^{1/2} C M_Q C_Q \left(1 + \|f\|^2 \right)^{1/2} \\ &= \rho^2 \left\| U_\rho(\mathbf{z})_{t-2} - \mathcal{U}_\rho^f(\mathbf{z})_{t-2} \right\| + (1 + \rho) \epsilon^{1/2} C M_Q C_Q \left(1 + \|f\|^2 \right)^{1/2} \\ &\leq \rho^j \left\| U_\rho(\mathbf{z})_{t-j} - \mathcal{U}_\rho^f(\mathbf{z})_{t-j} \right\| \\ &+ (1 + \rho + \rho^2 + \dots + \rho^{j-1}) \epsilon^{1/2} C M_Q C_Q \left(1 + \|f\|^2 \right)^{1/2}. \end{aligned} \quad (55)$$

As by hypothesis $\rho < 1$, we can take the limit $j \rightarrow \infty$ in this expression, which yields (26). In order to obtain (27) it suffices to replace the use of (20) in (54) by that of (22).

(v) First of all, note that for any $R > \max\{1/\|f\|^2, 1\}$, the contraction parameter $\rho = 1/(R\|f\|^2)$ satisfies the condition (24). Set now $Q := \{\pm \mathbf{e}_1, \dots, \pm \mathbf{e}_N\}$. It is easy to see that with this choice, the norm $\|\cdot\|_Q$ introduced in Lemma III.1 satisfies that $\|\cdot\|_Q = \|\cdot\|_1$ and

that $M_Q = 1$. If we now recall that $\|\cdot\| \leq \|\cdot\|_1 \leq \sqrt{N}\|\cdot\|$ and that $(1/\sqrt{N})\|\cdot\| \leq \|\cdot\|_1 \leq \sqrt{N}\|\cdot\|$, we can rewrite the inequality (54) as

$$\begin{aligned} & \left\| U_\rho(\mathbf{z})_t - \mathcal{U}_\rho^f(\mathbf{z})_t \right\| \leq \rho \left\| U_\rho(\mathbf{z})_{t-1} - \mathcal{U}_\rho^f(\mathbf{z})_{t-1} \right\| \\ &+ \epsilon^{1/2} \|\mathbb{I}_N - f^* \circ f\|_1^{1/2} \left\| F_\rho(\mathcal{U}_\rho^f(\mathbf{z})_{t-1}, \mathbf{z}_t) \right\|_1 \\ &\leq \rho \left\| U_\rho(\mathbf{z})_{t-1} - \mathcal{U}_\rho^f(\mathbf{z})_{t-1} \right\| + \epsilon^{1/2} N^{3/4} C \|\mathbb{I}_N - f^* \circ f\|_1^{1/2} \\ &\leq \epsilon^{1/2} N^{3/4} C \|\mathbb{I}_N - f^* \circ f\|_1^{1/2} \frac{1}{1 - \rho} \\ &\leq \epsilon^{1/2} N^{3/4} C (1 + \|f\|^2)^{1/2} \frac{R\|f\|^2}{R\|f\|^2 - 1}, \end{aligned}$$

where in the passage from the second to the third line we just used the same iterative bounding procedure as in (55). The bound in (29) is obtained by using the inequality in (23) adapted to our particular choice of Q , according to which, for any $\mathbf{v} \in \mathbb{R}^N$, we have that

$$\begin{aligned} & \|(\mathbb{I}_N - f^* \circ f)(\mathbf{v})\|^2 \leq \epsilon \|(\mathbb{I}_N - f^* \circ f)(\mathbf{v})\|_1 \|\mathbf{v}\|_1 \\ &\leq \epsilon \sqrt{N} \|(\mathbb{I}_N - f^* \circ f)(\mathbf{v})\| \|\mathbf{v}\|_1, \end{aligned}$$

which implies that

$$\|(\mathbb{I}_N - f^* \circ f)(\mathbf{v})\| \leq \epsilon \sqrt{N} \|\mathbf{v}\|_1. \quad (56)$$

Consequently, by (53) and (56), we have

$$\begin{aligned} & \left\| U_\rho(\mathbf{z})_t - \mathcal{U}_\rho^f(\mathbf{z})_t \right\| \leq \rho \left\| U_\rho(\mathbf{z})_{t-1} - \mathcal{U}_\rho^f(\mathbf{z})_{t-1} \right\| \\ &+ \left\| (\mathbb{I}_N - f^* \circ f)(F_\rho(\mathcal{U}_\rho^f(\mathbf{z})_{t-1}, \mathbf{z}_t)) \right\| \\ &\leq \rho \left\| U_\rho(\mathbf{z})_{t-1} - \mathcal{U}_\rho^f(\mathbf{z})_{t-1} \right\| + \epsilon \sqrt{N} \left\| F_\rho(\mathcal{U}_\rho^f(\mathbf{z})_{t-1}, \mathbf{z}_t) \right\|_1 \\ &\leq \epsilon N C \frac{1}{1 - \rho} = \epsilon N C \frac{R\|f\|^2}{R\|f\|^2 - 1}. \quad \blacksquare \end{aligned}$$

G. Proof of Corollary III.6

It consists of just applying the JL projection Theorem III.5 to the SigSAS system introduced in Proposition II.2. First of all, by (11), the state map $F_{\lambda, l, p}^{\text{SigSAS}} : \widetilde{B}_{\|\cdot\|}(\mathbf{0}, L) \times [-M, M] \rightarrow \widetilde{B}_{\|\cdot\|}(\mathbf{0}, L)$ is continuous and a contraction on the state variable, with contraction constant $\rho := \lambda \widetilde{M}$. Now, if we fix $R > \max\{1/\|f\|^2, 1/\widetilde{M}\|f\|^2, 1\}$, then $\lambda = 1/R\widetilde{M}\|f\|^2$ satisfies all the necessary constraints in the statement of Proposition II.2, as well as the condition $\rho < 1/\|f\|^2$, as required in (24).

Therefore, given a JL map f , part (i) of Theorem III.5 shows the existence of a contractive state map $F_{\lambda, l, p, f}^{\text{SigSAS}}$ with constant $\lambda \widetilde{M}\|f\|^2$ defined by $F_{\lambda, l, p, f}^{\text{SigSAS}}(\mathbf{x}, z) := f(F_{\lambda, l, p}^{\text{SigSAS}}(f^*(\mathbf{x}), z))$, $\mathbf{x} \in \mathbb{R}^k$, $z \in [-M, M]$. Following the same strategy as in (45), it is easy to see that $\left\| F_{\lambda, l, p, f}^{\text{SigSAS}}(\mathbf{0}, z) \right\| \leq \widetilde{M}\|f\|$, for all $z \in [-M, M]$, and hence by [54, Remark 2] we can conclude that $F_{\lambda, l, p, f}^{\text{SigSAS}}$ restricts to a map $F_{\lambda, l, p, f}^{\text{SigSAS}} : \widetilde{B}_{\|\cdot\|}(\mathbf{0}, L_f) \times [-M, M] \rightarrow \widetilde{B}_{\|\cdot\|}(\mathbf{0}, L_f)$, with $L_f = \widetilde{M}\|f\|/(1 - \lambda \widetilde{M}\|f\|^2)$. In those circumstances, part (iii) of Theorem III.5 allow us to conclude that the JL reduced SigSAS state map $\mathcal{F}_{\lambda, l, p, f}^{\text{SigSAS}} : f^*(\widetilde{B}_{\|\cdot\|}(\mathbf{0}, L_f)) \times [-M, M] \rightarrow f^*(\widetilde{B}_{\|\cdot\|}(\mathbf{0}, L_f))$ has the ESP and a unique FMP associated filter $\mathcal{U}_{\lambda, l, p, f}^{\text{SigSAS}} : K_M \rightarrow \left(f^*(\widetilde{B}_{\|\cdot\|}(\mathbf{0}, L_f))\right)^{\mathbb{Z}_-}$.

The error bounds in (30)-(31) are obtained out of (28)-(29). Indeed, we first notice that by (12) $\left\| F_{\lambda, l, p}^{\text{SigSAS}}(\mathbf{x}, z) \right\| \leq L$, with $L := \widetilde{M}/(1 - \lambda \widetilde{M})$, for any $\mathbf{x} \in \widetilde{B}_{\|\cdot\|}(\mathbf{0}, L)$ and $z \in [-M, M]$, and we can hence take L as the constant C in the bounds (28)-(29). Consequently, using the bound (28), we have

$$\begin{aligned} & \left\| W U_{\lambda, l, p}^{\text{SigSAS}}(\mathbf{z})_t - W \mathcal{U}_{\lambda, l, p, f}^{\text{SigSAS}}(\mathbf{z})_t \right\| \\ &\leq \|W\| \left\| U_{\lambda, l, p}^{\text{SigSAS}}(\mathbf{z})_t - \mathcal{U}_{\lambda, l, p, f}^{\text{SigSAS}}(\mathbf{z})_t \right\| \\ &\leq \|W\| \epsilon^{\frac{1}{2}} N^{\frac{3}{4}} (1 + \|f\|^2)^{\frac{1}{2}} \frac{R\|f\|^2 \widetilde{M}}{(R\|f\|^2 - 1)(1 - \lambda \widetilde{M})} \\ &= \|W\| \epsilon^{\frac{1}{2}} N^{\frac{3}{4}} (1 + \|f\|^2)^{\frac{1}{2}} \frac{\widetilde{M} R^2 \|f\|^4}{(R\|f\|^2 - 1)^2}. \end{aligned}$$

The second inequality follows from (29). $\blacksquare$

H. Proof of Theorem III.7

According to the third part of Theorem III.5, the JL reduction $\mathcal{F}_{\lambda,l,p,f}^{\text{SigSAS}}$ of $F_{\lambda,l,p}^{\text{SigSAS}}(\mathbf{x}, z) = \lambda \pi_l(\mathbf{x}) \otimes \tilde{\mathbf{z}} + \tilde{\mathbf{z}}^0$ is isomorphic to the map $F_{\lambda,l,p,f}^{\text{SigSAS}}$ given by

$$F_{\lambda,l,p,f}^{\text{SigSAS}}(\mathbf{x}, z) := \lambda f(\pi_l(f^*(\mathbf{x})) \otimes \tilde{\mathbf{z}}) + f(\tilde{\mathbf{z}}^0). \quad (57)$$

Take now the canonical bases in $T^{l+1}(\mathbb{R}^{p+1})$ (using the lexicographic order) and $\mathbb{R}^k$ and let $S \in \mathbb{M}_{k,N}$, $N = (p+1)^{l+1}$, be the random matrix associated to the JL projection f in those bases. Let also $N_0 := (p+1)^l$, we separate S in blocks and write:

$$S = (S_1 | \dots | S_{p+1}), \text{ with } S_1, \dots, S_{p+1} \in \mathbb{M}_{k,N_0}, S_{ij} \sim N(0, 1/k). \quad (58)$$

The order-lowering map $\pi_l : T^{l+1}(\mathbb{R}^{p+1}) \rightarrow T^l(\mathbb{R}^{p+1})$ takes in these bases the matrix form

$$\Pi_l = (\mathbb{I}_{N_0} | \mathbb{O}_{N_0,pN_0}) \quad \text{and} \quad \Pi_l S^\top = S_1^\top. \quad (59)$$

Consequently (57) can be rewritten in matrix form as

$$F_{\lambda,l,p,f}^{\text{SigSAS}}(\mathbf{x}, z) = \lambda \sum_{i=1}^{p+1} z^{i-1} S(\Pi_l(S^\top(\mathbf{x})) \otimes \mathbf{e}_i) + S\tilde{\mathbf{z}}^0. \quad (60)$$

We analyze this map by first spelling out the matrix associated to the linear assignment $\mathbf{x} \mapsto \Pi_l(S^\top(\mathbf{x})) \otimes \mathbf{e}_i$, $i = 1, \dots, p+1$. It is easy to check that:

$$\Pi_l(S^\top(\cdot)) \otimes \mathbf{e}_i =$$

$$\left(\begin{array}{cccc} \mathbb{O}_{i-1,1} & \mathbb{O}_{i-1,1} & \cdots & \mathbb{O}_{i-1,1} \\ S_{1,1} & S_{2,1} & \cdots & S_{k,1} \\ \mathbb{O}_{p-i+1,1} & \mathbb{O}_{p-i+1,1} & \cdots & \mathbb{O}_{p-i+1,1} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \mathbb{O}_{i-1,1} & \mathbb{O}_{i-1,1} & \cdots & \mathbb{O}_{i-1,1} \\ S_{1,(p+1)^l} & S_{2,(p+1)^l} & \cdots & S_{k,(p+1)^l} \\ \mathbb{O}_{p-i+1,1} & \mathbb{O}_{p-i+1,1} & \cdots & \mathbb{O}_{p-i+1,1} \end{array} \right) \begin{array}{l} \text{Block } 1 \\ \\ \\ \\ \\ \\ \text{Block } (p+1)^l \end{array}. \quad (61)$$

Using this matrix expression, it is easy to see that, for any $j, m \in \{1, \dots, k\}$:

$$\begin{aligned} (S(\Pi_l(S^\top(\cdot)) \otimes \mathbf{e}_i))_{j,m} &= \sum_{r=1}^N S_{j,r} (\Pi_l(S^\top(\cdot)) \otimes \mathbf{e}_i)_{r,m} \\ &= \sum_{n=1}^{N_0} S_{j,i+(n-1)(1+p)} S_{m,n}, \text{ with } N_0 = (p+1)^l, \end{aligned} \quad (62)$$

and hence by (58), each of these entries is the sum of the products of two independent zero mean normal random variables with variance $1/k$, unless those two factors are identical, which can only happen whenever $j = m$ and $i + (n-1)(1+p) = n$ simultaneously, which only holds for $i = 1, n = 1$, and for the diagonal terms $j = m$. This implies that

$$(S(\Pi_l(S^\top(\cdot)) \otimes \mathbf{e}_1))_{j,m} \sim \begin{cases} \sum_{n=1}^{N_0} a_{j,m,n}^1, & \text{when } j \neq m, \\ \sum_{n=1}^{N_0-1} a_{j,m,n}^1 + b_{j,m}, & \text{otherwise,} \end{cases} \quad (63)$$

where $b_{j,m} = (1/k)P_{j,m}$, $a_{j,m,n}^1 = (1/2k)(Q_{j,m,n} - R_{j,m,n})$, and $P_{j,m}, Q_{j,m,n}, R_{j,m,n} \sim \chi^2(1)$ are independent random variables. Analogously, for $i \in \{2, \dots, p+1\}$,

$$(S(\Pi_l(S^\top(\cdot)) \otimes \mathbf{e}_i))_{j,m} \sim \sum_{n=1}^{N_0} a_{j,m,n}^i, \quad (64)$$

with $a_{j,m,n}^i$ as above. We now study the matrix form of the summand $S\tilde{\mathbf{z}}^0$ in (60). First of all, by (34),

$$S\tilde{\mathbf{z}}^0 = rSC^{I_0}(1, z, \dots, z^p)^\top,$$

and hence it can be written as $B(1, z, \dots, z^p)^\top = S\tilde{\mathbf{z}}^0$, with $B \in \mathbb{M}_{k,p+1}$ the matrix with components

$$B_{j,m} = r \sum_{k=1}^N S_{jk} C_{km}^{I_0} = r S_{jm} \mathbf{1}_{\{m \in I_0\}}, \quad (65)$$

which (38), as the product of a Gaussian with a Rademacher random variable is Gaussian distributed.

We now prove the claim in (37) regarding the matrices $A_1, \dots, A_{p+1}$. First of all, (63) and (64) show that for each $i \in \{1, \dots, p+1\}$, the entries $(S(\Pi_l(S^\top(\cdot)) \otimes \mathbf{e}_i))_{j,m}$ of $\tilde{A}_i = \lambda_0 S(\Pi_l(S^\top(\cdot)) \otimes \mathbf{e}_i) = (\delta/2M)\sqrt{k/N_0} S(\Pi_l(S^\top(\cdot)) \otimes \mathbf{e}_i)$ may be of two types

$$(\tilde{A}_i)_{j,m} \sim \begin{cases} \frac{\delta\sqrt{k}}{2M} \sqrt{N_0} \left(\frac{\sum_{n=1}^{N_0-1} a_{j,m,n}^i}{N_0} + \frac{b_{j,m}}{N_0} \right), & \text{if } i=1, j=m, \\ \frac{\delta\sqrt{k}}{2M} \sqrt{N_0} \left(\frac{\sum_{n=1}^{N_0} a_{j,m,n}^i}{N_0} \right), & \text{otherwise.} \end{cases} \quad (66)$$

As long as we are in the cases $j \neq m$ or simultaneously $i = 1$ and $j = m$, these entries are the sum of IID mean zero random variables of variance $1/k^2$ and hence by the Lindeberg Central Limit Theorem, they converge in distribution to mean zero Gaussian random variables $(A_i)_{j,m} \sim N\left(0, \frac{\delta^2}{4kM^2}\right)$, as required.

This straightforward argument cannot be used when $j = m$ and $i \geq 2$ as, in that situation, the random variable $S_{m,i}$ appears in two different summands in the expression (62). The claim in that case is proved by using a martingale central limit theorem. Indeed, consider the right hand side of the equality (62) for fixed $i \geq 2$ and $j = m$, that is,

$$\sum_{n=1}^{N_0} S_{j,i+(n-1)(1+p)} S_{j,n}. \quad (67)$$

As the entries $S_{j,n}$ are independent zero mean normal random variables with variance $1/k$, we have,

$$\text{Var} \left(\sum_{n=1}^{N_0} S_{j,i+(n-1)(1+p)} S_{j,n} \right) = \frac{N_0}{k^2}. \quad (68)$$

Notice first that for $s \geq l$, we have that

$$\mathbb{E} \left[\frac{k}{\sqrt{N_0}} \sum_{n=1}^s S_{j,i+(n-1)(1+p)} S_{j,n} \middle| \mathcal{F}_l \right] = \frac{k}{\sqrt{N_0}} \sum_{n=1}^l S_{j,i+(n-1)(1+p)} S_{j,n},$$

where $\mathcal{F}_l$ is the σ -algebra generated by $\{S_{j,i+(n-1)(1+p)}, S_{j,n} \mid n \leq l\}$. This equality is a consequence of the fact that for $s \geq l$ at most one of the terms in the product

$$S_{j,i+(s-1)(1+p)} S_{j,s}$$

is $\mathcal{F}_l$ -measurable and the other one is independent with zero expectation. We shall apply Theorem 3.2 in [56] using as martingale differences the random variables $Y_{N_0,n}$ defined by

$$Y_{N_0,n} := \frac{k}{\sqrt{N_0}} S_{j,i+(n-1)(1+p)} S_{j,n}.$$

It is clear that $\max_n \{Y_{N_0,n}\} \rightarrow 0$ in probability as N_0 tends to ∞ .

Moreover, $\mathbb{E} \left[\max_n \{Y_{N_0,n}^2\} \right]$ is bounded in N_0 and

$$\lim_{N_0 \rightarrow \infty} \sum_{n=1}^{N_0} Y_{N_0,n}^2 = 1.$$

Indeed,

$$\begin{aligned} \mathbb{E} \left[\left(\sum_{n=1}^{N_0} Y_{N_0,n}^2 - 1 \right)^2 \right] &= 2 \sum_{s < n}^{N_0} \mathbb{E} \left[\left(Y_{N_0,s}^2 - \frac{1}{N_0} \right) \left(Y_{N_0,n}^2 - \frac{1}{N_0} \right) \right] \\ &\quad + \sum_{n=1}^{N_0} \mathbb{E} \left[\left(Y_{N_0,n}^2 - \frac{1}{N_0} \right)^2 \right], \end{aligned} \quad (69)$$

where for the second sum, using $\sum_{n=1}^{N_0} \mathbb{E} [Y_{N_0,n}^4] = N_0 \frac{k^4}{N_0^2} \frac{9}{k^4} = 9/N_0$ and that $Y_{N_0,n}$, $n = 1, \dots, N_0$, are random variables with zero mean and variance $1/N_0$, it holds that

$$\sum_{n=1}^{N_0} \mathbb{E} \left[\left(Y_{N_0,n}^2 - \frac{1}{N_0} \right)^2 \right] = 8/N_0 \xrightarrow{N_0 \rightarrow \infty} 0.$$

Notice now that in the first sum in (69) the terms involving independent martingale differences vanish. Hence, it is only the summands which correspond to the case when $Y_{N_0,s}$ and $Y_{N_0,n}$ are made of non-independent random variables that need to be treated separately. We notice that this happens precisely when $i + (s-1)(1+p) = n$ and write

$$\begin{aligned} & 2 \sum_{s < n}^{N_0} \mathbb{E} \left[\left(Y_{N_0,s}^2 - \frac{1}{N_0} \right) \left(Y_{N_0,n}^2 - \frac{1}{N_0} \right) \right] \\ &= 2 \sum_{s < n}^{N_0} \mathbb{E} \left[Y_{N_0,s}^2 Y_{N_0,n}^2 - \frac{1}{N_0^2} \right] \\ &= 2 \sum_{s > 1 + \frac{1-i}{p}}^{N_0} \mathbb{E} \left[\frac{k^4}{N_0^2} S_{j,i+(s-1)(1+p)}^4 S_{j,s}^2 S_{j,i+(i-1)(1+p)+(s-1)(1+p)}^2 \right. \\ &\quad \left. - \frac{1}{N_0^2} \right] \leq 2(N_0 - 1) \left(\frac{k^4}{N_0^2} \frac{3}{k^2} \frac{1}{k^2} - \frac{1}{N_0^2} \right) = 4(N_0 - 1)/N_0^2 \xrightarrow{N_0 \rightarrow \infty} 0. \end{aligned}$$

Hence by Theorem 3.2 in [56] $Y_{N_0,n}$ converges in distribution to a standard Gaussian random variable and the claim follows.

The mutual independence between the components of the different matrices A_i and within each individual matrix is obtained by using a multivariate version of the martingale central limit theorem: the arguments are fully in line with the previous application of the martingale central limit theorem. The same applies for the statement about B .

We conclude by showing that with the choice of λ_0 in (35), the reduced SAS system $F_{\lambda_0,l,p,f}^{\text{SigSAS}}$ in (36) has the ESP and the FMP with probability at least $1 - \delta$. Let $p(z) := \sum_{i=1}^{p+1} z^{i-1} A_i$ and recall that by [43, Proposition 14], the system $F_{\lambda_0,l,p,f}^{\text{SigSAS}}$ has the ESP and the FMP whenever

$$M_p = \sup_{z \in [-M,M]} \{ \|p(z)\| \} < 1.$$

Given that

$$M_p = \sup_{z \in [-M,M]} \left\{ \left\| \sum_{i=1}^{p+1} z^{i-1} A_i \right\| \right\} \leq \sum_{i=1}^{p+1} M^{i-1} \|A_i\|,$$

it is clear that the following inclusion of events holds $\{M_p \geq 1\} \subset \left\{ \sum_{i=1}^{p+1} M^{i-1} \|A_i\| \geq 1 \right\}$. This implies that

$$\begin{aligned} \mathbb{P}(M_p \geq 1) &\leq \mathbb{P} \left(\sum_{i=1}^{p+1} M^{i-1} \|A_i\| \geq 1 \right) \\ &\leq \sum_{i=1}^{p+1} M^{i-1} \mathbb{E} [\|A_i\|] \leq \widetilde{M} 2\sqrt{k} \frac{\delta}{2\sqrt{k}\widetilde{M}} = \delta, \end{aligned}$$

as required. Note that in the second inequality we used Markov's inequality and that the third one is a consequence of Gordon's Theorem (32) and of (37). ■

I. Proof of Theorem III.8

First of all, given $l, p \in \mathbb{N}$, consider the SigSAS filter $U_{\lambda_0,l,p}^{\text{SigSAS}}$ associated to the parameter λ_0 defined in (35) and the linear map $\overline{W} \in L(T^{l+1}(\mathbb{R}^{p+1}), \mathbb{R}^m)$ whose existence for the approximation of U is guaranteed by Theorem II.4. Consider now a JL projection $f: \mathbb{R}^N \rightarrow \mathbb{R}^k$ that satisfies (17) with a parameter $0 < \epsilon < 1$ that satisfies (16) with n replaced by N and the value $k \in \mathbb{N}$ in the statement. Let now $F_{\lambda_0,l,p,f}^{\text{SigSAS}}$ be the JL reduced SigSAS system given by Corollary III.6 and that asymptotically takes the form (39) because of (36) in Theorem III.7. Let $W \in L(T^{l+1}(\mathbb{R}^{p+1}), \mathbb{R}^m)$ be the linear map in (15) in Theorem II.4. Define $\mathcal{W} := W \circ i_k$ as in Corollary III.6

and $\overline{W} := \mathcal{W} \circ f^* = W \circ f^*$. We now use the triangle inequality and the third part of Theorem III.5 to write:

$$\begin{aligned} & \left\| U(\mathbf{z})_t - \overline{W} U_{\lambda_0,l,p,f}^{\text{SigSAS}}(\mathbf{z})_t \right\| \\ &\leq \left\| U(\mathbf{z})_t - W U_{\lambda_0,l,p}^{\text{SigSAS}}(\mathbf{z})_t \right\| + \left\| W U_{\lambda_0,l,p}^{\text{SigSAS}}(\mathbf{z})_t - \overline{W} U_{\lambda_0,l,p,f}^{\text{SigSAS}}(\mathbf{z})_t \right\| \\ &\leq \left\| U(\mathbf{z})_t - W U_{\lambda_0,l,p}^{\text{SigSAS}}(\mathbf{z})_t \right\| + \left\| W U_{\lambda_0,l,p}^{\text{SigSAS}}(\mathbf{z})_t - \mathcal{W} U_{\lambda_0,l,p,f}^{\text{SigSAS}}(\mathbf{z})_t \right\| \\ &\leq w_l^U + L \left(1 - \frac{\|\mathbf{z}\|_\infty}{M} \right)^{-1} \left(\frac{\|\mathbf{z}\|_\infty}{M} \right)^{p+1} + I_{l,p}. \end{aligned}$$

In the last inequality we have used Theorem II.4 to bound the first summand and the last bounds stated in Corollary III.6 for the second one. More specifically, regarding the last point, we first note that the parameter R in (30) and (31) is determined in this case by the definition of λ_0 in (35) and the equality

$$\frac{\delta}{2\widetilde{M}} \sqrt{\frac{k}{N_0}} = \frac{1}{R\widetilde{M}\|f\|^2}$$

with $N_0 = (p+1)^l$. Hence, the bounds in (30) and (31) can be written as

$$\begin{aligned} & \|W\| \epsilon^{\frac{1}{2}} N^{\frac{3}{4}} (1 + \|f\|^2)^{\frac{1}{2}} \frac{\widetilde{M} R^2 \|f\|^4}{(R\|f\|^2 - 1)^2} \\ &= \|W\| \epsilon^{\frac{1}{2}} N^{\frac{3}{4}} \widetilde{M} \frac{(1 + \|f\|^2)^{\frac{1}{2}}}{\left(1 - \frac{\delta}{2} \sqrt{\frac{k}{N_0}}\right)^2} \quad \text{and} \\ & \|W\| \epsilon N \frac{\widetilde{M} R^2 \|f\|^4}{(R\|f\|^2 - 1)^2} = \|W\| \epsilon N \widetilde{M} \frac{1}{\left(1 - \frac{\delta}{2} \sqrt{\frac{k}{N_0}}\right)^2}, \end{aligned}$$

respectively, which prove (41). ■

ACKNOWLEDGMENT

CC acknowledges partial financial support from the Vienna Science and Technology Fund (WWTF) grant MA16-021, FWF START Grant Y 1235. LGonon and JPO acknowledge partial financial support coming from the Research Commission of the Universität Sankt Gallen and the Swiss National Science Foundation (grant number 200021_175801/1). JPO acknowledges partial financial support of the French ANR “BIPHOPROC” project (ANR-14-OHRI-0002-02). JT acknowledges support from the ETH Foundation and the Swiss National Science Foundation (grant number 179114, “Machine Learning in Finance”). The authors thank the hospitality and the generosity of the FIM at ETH Zurich and the Division of Mathematical Sciences of the Nanyang Technological University, Singapore, where a significant portion of the results in this paper were obtained.

REFERENCES

- [1] W. Maass and E. D. Sontag, “Neural Systems as Nonlinear Filters,” *Neural Computation*, vol. 12, no. 8, pp. 1743–1772, aug 2000. [Online]. Available: <http://www.mitpressjournals.org/doi/abs/10.1162/089976600300015123>
- [2] W. Maass, T. Natschlager, 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.
- [3] T. Natschlager, W. Maass, and H. Markram, “The “Liquid Computer”: a novel strategy for real-time computing on time series,” *Special Issue on Foundations of Information Processing of TELEMATIK*, vol. 8, no. 1, pp. 39–43, 2002. [Online]. Available: <http://infoscience.epfl.ch/record/117806>

- [4] W. Maass, T. Natschlager, 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.
- [5] 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. [Online]. Available: <http://dx.plos.org/10.1371/journal.pcbi.0020165>
- [6] 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: <http://www.research-collection.ethz.ch/443/handle/20.500.11850/140592>
- [7] M. Matthews and G. Moschytz, "The identification of nonlinear discrete-time fading-memory systems using neural network models," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 41, no. 11, pp. 740–751, 1994. [Online]. Available: <http://ieeexplore.ieee.org/document/331544/>
- [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] M. Lukoševičius and H. Jaeger, "Reservoir computing approaches to recurrent neural network training," *Computer Science Review*, vol. 3, no. 3, pp. 127–149, 2009.
- [10] L. Grigoryeva, J. Henriques, L. Larger, and J.-P. Ortega, "Stochastic time series forecasting using time-delay reservoir computers: performance and universality," *Neural Networks*, vol. 55, pp. 59–71, 2014.
- [11] A. Goudarzi, S. Marzen, P. Banda, G. Feldman, M. R. Lakin, C. Teuscher, and D. Stefanovic, "Memory and information processing in recurrent neural networks," Tech. Rep., 2016. [Online]. Available: <https://arxiv.org/pdf/1604.06929.pdf>
- [12] S. Marzen, "Difference between memory and prediction in linear recurrent networks," *Physical Review E*, vol. 96, no. 3, pp. 1–7, 2017.
- [13] J. Pathak, Z. Lu, B. R. Hunt, M. Girvan, and E. Ott, "Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data," *Chaos*, vol. 27, no. 12, 2017.
- [14] J. Pathak, B. Hunt, M. Girvan, Z. Lu, and E. Ott, "Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data: A Reservoir Computing Approach," *Physical Review Letters*, vol. 120, no. 2, p. 24102, 2018. [Online]. Available: <https://doi.org/10.1103/PhysRevLett.120.024102>
- [15] Z. Lu, B. R. Hunt, and E. Ott, "Attractor reconstruction by machine learning," *Chaos*, vol. 28, no. 6, 2018.
- [16] L. Appeltant, M. C. Soriano, G. Van der Sande, J. Danckaert, S. Massar, J. Dambre, B. Schrauwen, C. R. Mirasso, and I. Fischer, "Information processing using a single dynamical node as complex system," *Nature Communications*, vol. 2, p. 468, jan 2011.
- [17] A. Rodan and P. Tino, "Minimum complexity echo state network," *IEEE Transactions on Neural Networks*, vol. 22, no. 1, pp. 131–44, jan 2011.
- [18] K. Vandoorne, J. Dambre, D. Verstraeten, B. Schrauwen, and P. Bienstman, "Parallel reservoir computing using optical amplifiers," *IEEE Transactions on Neural Networks*, vol. 22, no. 9, pp. 1469–1481, sep 2011. [Online]. Available: <http://ieeexplore.ieee.org/document/5966352/>
- [19] L. Larger, M. C. Soriano, D. Brunner, L. Appeltant, J. M. Gutierrez, L. Pesquera, C. R. Mirasso, and I. Fischer, "Photonic information processing beyond Turing: an optoelectronic implementation of reservoir computing," *Optics Express*, vol. 20, no. 3, p. 3241, jan 2012.
- [20] Y. Paquot, F. Duport, A. Smerieri, J. Dambre, B. Schrauwen, M. Haelterman, and S. Massar, "Optoelectronic reservoir computing," *Scientific reports*, vol. 2, p. 287, jan 2012.
- [21] D. Brunner, M. C. Soriano, C. R. Mirasso, and I. Fischer, "Parallel photonic information processing at gigabyte per second data rates using transient states," *Nature Communications*, vol. 4, no. 1364, 2013.
- [22] K. Vandoorne, P. Mechet, T. Van Vaerenbergh, M. Fiers, G. Morthier, D. Verstraeten, B. Schrauwen, J. Dambre, and P. Bienstman, "Experimental demonstration of reservoir computing on a silicon photonics chip," *Nature Communications*, vol. 5, pp. 78–80, mar 2014. [Online]. Available: <http://www.nature.com/doi/10.1038/ncomms4541>
- [23] Q. Vinckier, F. Duport, A. Smerieri, K. Vandoorne, P. Bienstman, M. Haelterman, and S. Massar, "High-performance photonic reservoir computer based on a coherently driven passive cavity," *Optica*, vol. 2, no. 5, pp. 438–446, 2015. [Online]. Available: <https://www.osapublishing.org/abstract.cfm?uri=optica-2-5-438>
- [24] F. Laporte, A. Katumba, J. Dambre, and P. Bienstman, "Numerical demonstration of neuromorphic computing with photonic crystal cavities," *Optics Express*, vol. 26, no. 7, p. 7955, apr 2018.
- [Online]. Available: <https://www.osapublishing.org/abstract.cfm?URI=oe-26-7-7955>
- [25] K. Doya, "Bifurcations in the learning of recurrent neural networks," in *Proceedings of IEEE International Symposium on Circuits and Systems*, vol. 6. IEEE, 1992, pp. 2777–2780. [Online]. Available: <http://ieeexplore.ieee.org/document/230622/>
- [26] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, may 2013, pp. 6645–6649. [Online]. Available: <http://ieeexplore.ieee.org/document/6638947/>
- [27] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, "How to construct deep recurrent neural networks," *arXiv*, dec 2013. [Online]. Available: <http://arxiv.org/abs/1312.6026>
- [28] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," *arXiv*, sep 2014. [Online]. Available: <http://arxiv.org/abs/1409.2329>
- [29] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," *Advances in neural information*, 2007.
- [30] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, dec 2006.
- [31] A. G. Hart, J. L. Hook, and J. H. P. Dawes, "Embedding and approximation theorems for echo state networks," *Neural Networks*, vol. 128, pp. 234–247, 2020. [Online]. Available: <http://arxiv.org/abs/1908.05202>
- [32] L. Gonon, L. Grigoryeva, and J.-P. Ortega, "Approximation error estimates for random neural networks and reservoir systems," *arXiv preprint 2002.05933*, 2020.
- [33] L. Grigoryeva and J.-P. Ortega, "Universal discrete-time reservoir computers with stochastic inputs and linear readouts using non-homogeneous state-affine systems," *Journal of Machine Learning Research*, vol. 19, no. 24, pp. 1–40, 2018. [Online]. Available: <http://arxiv.org/abs/1712.00754>
- [34] L. Gonon and J.-P. Ortega, "Reservoir computing universality with stochastic inputs," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 1, pp. 100–112, 2020.
- [35] K.-T. Chen, "Integration of paths, geometric invariants and a generalized Baker-Hausdorff formula," *Annals of Mathematics*, pp. 163–178, 1957.
- [36] —, "Integration of paths—A faithful representation of paths by non-commutative formal power series," *Transactions of the American Mathematical Society*, vol. 89, no. 2, pp. 395–407, 1958.
- [37] T. Lyons, "Differential equations driven by rough signals," *Revista Matemática Iberoamericana*, vol. 14, no. 2, pp. 215–310, 1998.
- [38] B. Hambly and T. Lyons, "Uniqueness for the signature of a path of bounded variation and the reduced path group," *Annals of Mathematics*, pp. 109–167, 2010.
- [39] C. Cuchiero, L. Gonon, L. Grigoryeva, J.-P. Ortega, and J. Teichmann, "Approximation of dynamics by randomized signature," 2019.
- [40] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," *Contemporary Mathematics*, vol. 26, pp. 189–206, 1984.
- [41] Y. C. Eldar, G. Kutyniok, and (editors), *Compressed Sensing*. Cambridge University Press, 2012.
- [42] L. Grigoryeva and J.-P. Ortega, "Echo state networks are universal," *Neural Networks*, vol. 108, pp. 495–508, 2018.
- [43] —, "Differentiable reservoir computing," *Journal of Machine Learning Research*, vol. 20, no. 179, pp. 1–62, 2019.
- [44] P. Lancaster and H. K. Farahat, "Norms on direct sums and tensor products," *Mathematics of Computation*, vol. 26, no. 118, pp. 401–401, may 1972. [Online]. Available: <http://www.ams.org/jourcgi/jour-getitem?pii=S0025-5718-1972-0305099-X>
- [45] C. Xu, "Hankel tensors, Vandermonde tensors and their positivities," *Linear Algebra and Its Applications*, vol. 491, pp. 56–72, 2016. [Online]. Available: <http://dx.doi.org/10.1016/j.laa.2015.02.012>
- [46] 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.
- [47] P. Frankl and H. Maehara, "The Johnson-Lindenstrauss lemma and the sphericity of some graphs," *Journal of Combinatorial Theory, Series B*, vol. 44, no. 3, pp. 355–362, 1988.
- [48] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, 1998, pp. 604–613.
- [49] S. Dasgupta and A. Gupta, "An elementary proof of a theorem of Johnson and Lindenstrauss," *Random Structures & Algorithms*, vol. 22, no. 1, pp. 60–65, 2003.

- [50] D. Achlioptas, “Database-friendly random projections: Johnson-Lindenstrauss with binary coins,” *Journal of computer and System Sciences*, vol. 66, no. 4, pp. 671–687, 2003.
- [51] A. Dasgupta, R. Kumar, and T. Sarlós, “A sparse Johnson-Lindenstrauss transform,” in *Proceedings of the forty-second ACM symposium on Theory of computing*, 2010, pp. 341–350.
- [52] J. Matoušek, “On variants of the Johnson-Lindenstrauss lemma,” *Random Structures & Algorithms*, vol. 33, no. 2, pp. 142–156, 2008.
- [53] R. Vershynin, “Introduction to the non-asymptotic analysis of random matrices,” in *Compressed Sensing*. Cambridge University Press, 2012, pp. 210–268.
- [54] L. Gonon, L. Grigoryeva, and J.-P. Ortega, “Risk bounds for reservoir computing,” *Preprint*, 2019.
- [55] —, “Memory and forecasting capacities of nonlinear recurrent networks,” *To appear in Physica D*, 2020.
- [56] P. Hall and C. C. Heyde, *Martingale Limit Theory and Its Application*. Academic Press, 1980.



Juan-Pablo Ortega holds a PhD in Mathematics from the University of California, Santa Cruz. He is a professor at the Faculty of Mathematics and Statistics at the University of St. Gallen (Switzerland) and a senior scientist at the French Centre National de la Recherche Scientifique (CNRS) (currently on leave). Juan-Pablo started his academic career working on geometric mechanics. His current interests are at the interface between geometry, dynamics, and machine learning.



Christa Cuchiero is a professor at the Institute of Statistics and Operations Research at the University of Vienna, Austria. She obtained her Ph.D. in Mathematics at ETH Zürich in 2011. Her research interests include mathematical finance, stochastic analysis, machine learning in finance, as well as statistics of stochastic processes.



Lukas Gonon is an assistant professor at the Department of Mathematics at the Ludwig-Maximilians-Universität München, Germany. He obtained his doctoral degree in mathematics at ETH Zürich, Switzerland. Subsequently, he worked as a post-doctoral researcher at the Faculty of Mathematics and Statistics, University of St. Gallen, Switzerland. His research interests are at the interface of mathematics, machine learning, and finance.



Lyudmila Grigoryeva is an assistant professor in computational statistics at the University of Konstanz, Germany. She obtained her PhD in mathematical modeling and computational methods at the Taras Shevchenko National University of Kyiv, Ukraine. Her research interests lie in the areas of statistical modeling, dynamical systems and stability, machine/statistical learning for dynamic processes, in particular reservoir computing.



Josef Teichmann has a PhD from Vienna University in analysis on Lie groups in 1999. He is a Professor for Mathematical Finance at ETH Zurich since 2009. He works in Mathematical Finance, Stochastic Partial Differential Equations, Rough Analysis, and Machine Learning techniques in Finance.