The universality problem in dynamic machine learning with applications to realized covolatilities forecasting

Lukas Gonon[1,2], Lyudmila Grigoryeva[3], and **Juan-Pablo Ortega**[2,4]

[1]ETH Zürich, Switzerland

[2]Universität Sankt Gallen, Switzerland

[3]Universität Konstanz, Germany

[4]CNRS, France

Innovative Research in Mathematical Finance
CIRM, Luminy
September 2018

University of St.Gallen

Outline

1. The universality problem in machine learning
2. Example with neural networks: famous approximation theorems
3. Dynamic problems and reservoir computing
4. Dynamic universality for deterministic inputs via Stone-Weierstrass
5. Extension to stochastic inputs
6. Dynamic universality via internal approximation: echo state networks are universal

University of St.Gallen

Publications

- Grigoryeva, L. and Ortega, J.-P. [2018] Universal discrete-time reservoir computers with stochastic inputs and linear readouts using non-homogeneours state-affine systems. *Journal of Machine Learning Research*, 19(24), 1-40.
- Grigoryeva, L. and Ortega, J.-P. [2018] Echo state networks are universal. To appear in *Neural Networks*.
- Gonon, L. and Ortega, J.-P. [2018] Reservoir computing universality with stochastic inputs. Preprint.

University of St.Gallen

Setup

- Machine learning as an input/output problem:
  - **Input z** contains available information for the solution of the problem (historical data, explanatory factors, features of the individuals that need to be classified).
  - **Output y** contains the solution of the problem (forecasted data, explained variables, classification results).
- Problem consists in determining (learning) **function(al)s** from **z** to **y**.
- We distinguish between static, dynamic, discrete-time, and continuous-time setups and between deterministic and stochastic situations.

The **universality problem** refers generically to the characterization of the space of **function(al)s** from **z** to **y** that can be learnt.

University of St.Gallen

Setups considered

	Static		Dynamic (discrete time)	
	Deterministic	**Stochastic**	**Deterministic**	**Stochastic**
Ingredients	$\mathbf{z} \in \mathbb{R}^n$ $\mathbf{y} \in \mathbb{R}^d$	$\mathbf{z} \in L^p(\Omega, \mathbb{R}^n)$ $\mathbf{y} \in L^p(\Omega, \mathbb{R}^d)$	$\mathbf{z} \in (\mathbb{R}^n)^{\mathbb{Z}_-}$ $\mathbf{y} \in (\mathbb{R}^d)^{\mathbb{Z}_-}$	$\mathbf{z} \in L^p\left(\Omega, (\mathbb{R}^n)^{\mathbb{Z}_-}\right)$ $\mathbf{y} \in L^p\left(\Omega, (\mathbb{R}^d)^{\mathbb{Z}_-}\right)$
Problem to be solved	$\mathbf{y} = f(\mathbf{z})$ f measurable	$\mathrm{E}[\mathbf{y} \mid \mathbf{z}]$	$\mathbf{y}(\cdot) = F(\mathbf{z}(\cdot))$	$\mathrm{E}[\mathbf{y}(\cdot) \mid \mathbf{z}(\cdot)]$
Examples and Applications	• observables or diagnostics variables in complex physical or noiseless engineering systems • translators • transcription	• image classification • speech recognition • factor analysis • anomaly detection	• integration or path continuation of (chaotic) differential equations • molecular dynamics • structural mechanics • vibration analysis • space mission design • autopilot systems • robotics • memory tasks • games	• physiological time series classification • financial bubble detection • time series forecasting • volatility filtering • system identification (blackboxing) • filters (transducers) and equalizers • imputation of missing values • source separators

University of St.Gallen

Discrete vs continuous-time

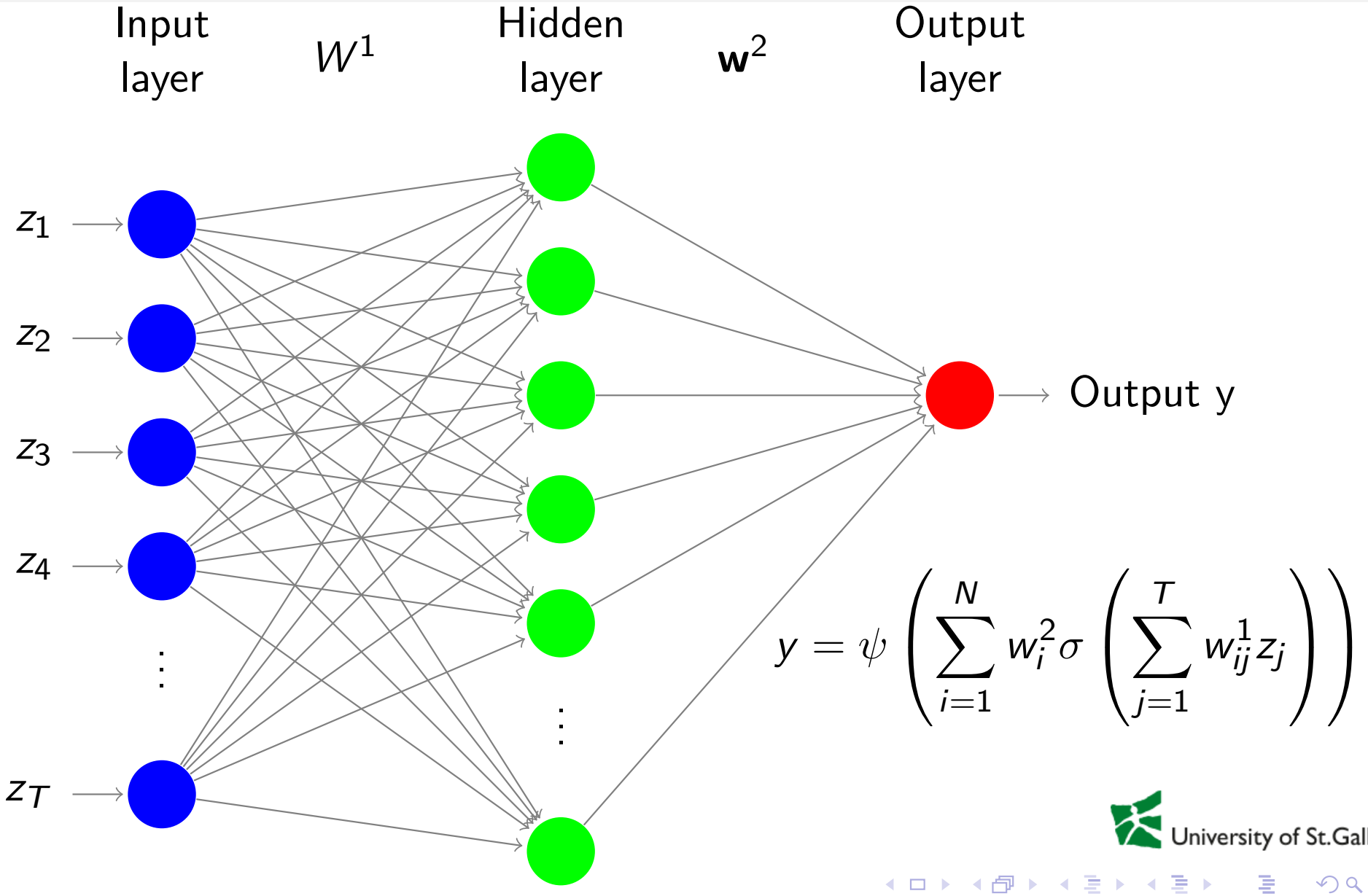
The discrete-time setup is technically much more complicated.

"...il convient de rappeler que les systèmes non linéaires en temps discret sont beaucoup plus mystérieux que les continus. Pour ces derniers, un certain nombre de techniques relevant de l'analyse fonctionnelle, de la géométrie différentielle ou des variables non commutatives sont disponibles. En discret, il n'y a rien ou presque. Or, il faut souligner que l'informatique tend à privilégier les systèmes discrets."

Michel Fliess and Dorothée Normand-Cyrot (1976)

University of St.Gallen

The neural networks example



University of St.Gallen

Universality in neural networks and approximation theorems

- Implemented as a machine learning device by tuning the weights $\mathbf{w}^i$ using a gradient descent algorithm (backpropagation) that minimizes the approximation error based on a training set.
- **Universality problem:** how large is the class of input-output functions that can be generated using feedforward neural networks?

University of St.Gallen

The Kolmogorov-Arnold-Sprecher representation theorems

Theorem (Kolmogorov-Arnold [Kol56, Arn57, Spr65, Spr96, Spr97])

There exist constants λ_p and fixed continuous increasing functions $\varphi_q(x)$ on $I = [0,1]$ such that each continuous function f on I^n can be written as

$$f(x_1, \ldots, x_n) = \sum_{q=1}^{2n+1} g_q \left(\sum_{p=1}^{n} \lambda_p \varphi_q(x_p) \right)$$

where the g_q are properly chosen continuous functions of one variable.

- The only genuinely multivariate function is the sum!
- This is a representation and not an approximation theorem
- The g_q functions depend on f but not λ_p and φ_q.
- Not ideal for machine learning applications: need to train the g_q.
- Extended to measurable functions: Rüschendorf and Thomsen [RT98].

The Cybenko and the Hornik *et al.* theorems

Theorem (Cybenko [Cyb89], Hornik, Stinchcombe, and White [HSW89])

Let σ be a continuous squashing function. Then, the functions $G_{\sigma,N} : I^n \to \mathbb{R}$ of the form

$$G_{\sigma,N}(\mathbf{z};\boldsymbol{\theta}) = \left(\sum_{1=1}^{N} w_j^2 \sigma\left(\langle \mathbf{w}_j^1, \mathbf{z} \rangle + \theta_j \right) \right), \quad \mathbf{w}_j^1, \mathbf{z} \in \mathbb{R}^n, \mathbf{w}^2 \in \mathbb{R}^N, \theta_j \in \mathbb{R},$$

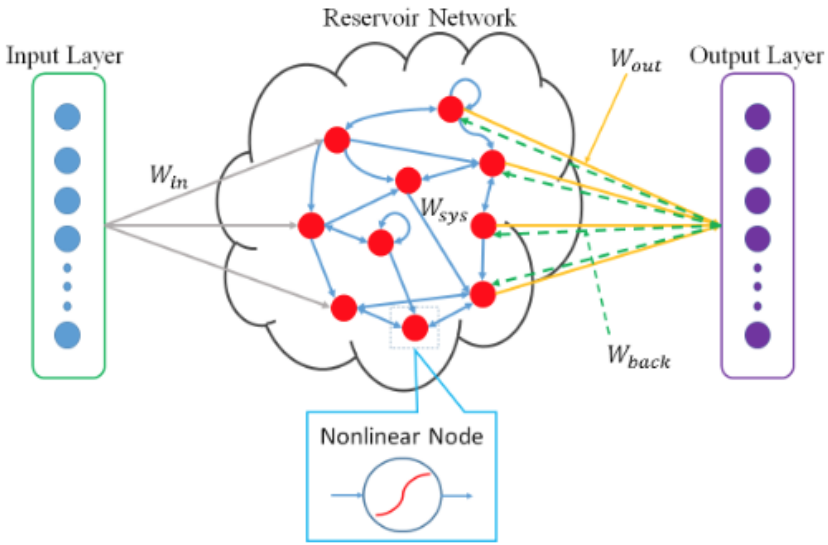
are dense in $C(I^n)$, that is, given any function $f \in C(I^n)$ and $\epsilon > 0$, there is a sum of this type for which

$$|G_{\sigma,N}(\mathbf{z};\boldsymbol{\theta}) - f(\mathbf{z})| < \epsilon, \quad \text{for all} \quad \mathbf{z} \in I^n.$$

- This result proves that any continuous function can be **approximated** using a feedforward neural network with a single hidden layer and continuous activation function.

University of St.Gallen

Dynamic problems and reservoir computing



- Modification of traditional RNN in which the architecture and the neuron weights of the network are created in advance (for example randomly) and remain unchanged during the training stage
- Compatible with high performance hardware implementations
- If readout layer is linear:
  - Data intensive applications become tractable
  - Inference and theoretical performance evaluation becomes possible!!

University of St.Gallen

Mathematical formulation of reservoir computing

A **reservoir computer (RC)** is a type of recurrent neural network (RNN):

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

determined by a **reservoir** map $F : \mathbb{R}^N \times \mathbb{R}^n \longrightarrow \mathbb{R}^N$ and a **readout** map $h : \mathbb{R}^N \to \mathbb{R}$ that transform (or filter) an infinite discrete-time input $\mathbf{z} = (\ldots, \mathbf{z}_{-1}, \mathbf{z}_0, \mathbf{z}_1, \ldots) \in (\mathbb{R}^n)^{\mathbb{Z}}$ into an output signal $\mathbf{y} \in \mathbb{R}^{\mathbb{Z}}$.

- $\mathbf{z}_t \in \mathbb{R}^n$ is an input signal, $\mathbf{x}_t \in \mathbb{R}^N$ is the **reservoir state**.
- The static readout $h : \mathbb{R}^N \to \mathbb{R}$ is trained in order to obtain the desired output y_t out of the input $\mathbf{z}_t$.
- **Multitasking:** different readouts can be trained on the same reservoir output to extract different pieces of information about the input.

University of St.Gallen

Forecasting of a Mackey-Glass chaotic time series

- We take one solution of the TDDE:

$$\frac{dx}{dt} = \frac{0.2x(t-\tau)}{1+x(t-\tau)^{10}-0.1x(t)} \quad \text{delay} \quad \tau = 17.$$

- We forecast a chaotic path by learning not the forecasting functional but by learning the dynamical system.

University of St.Gallen

We forecast with a very simple reservoir computer called **Echo State Network**:

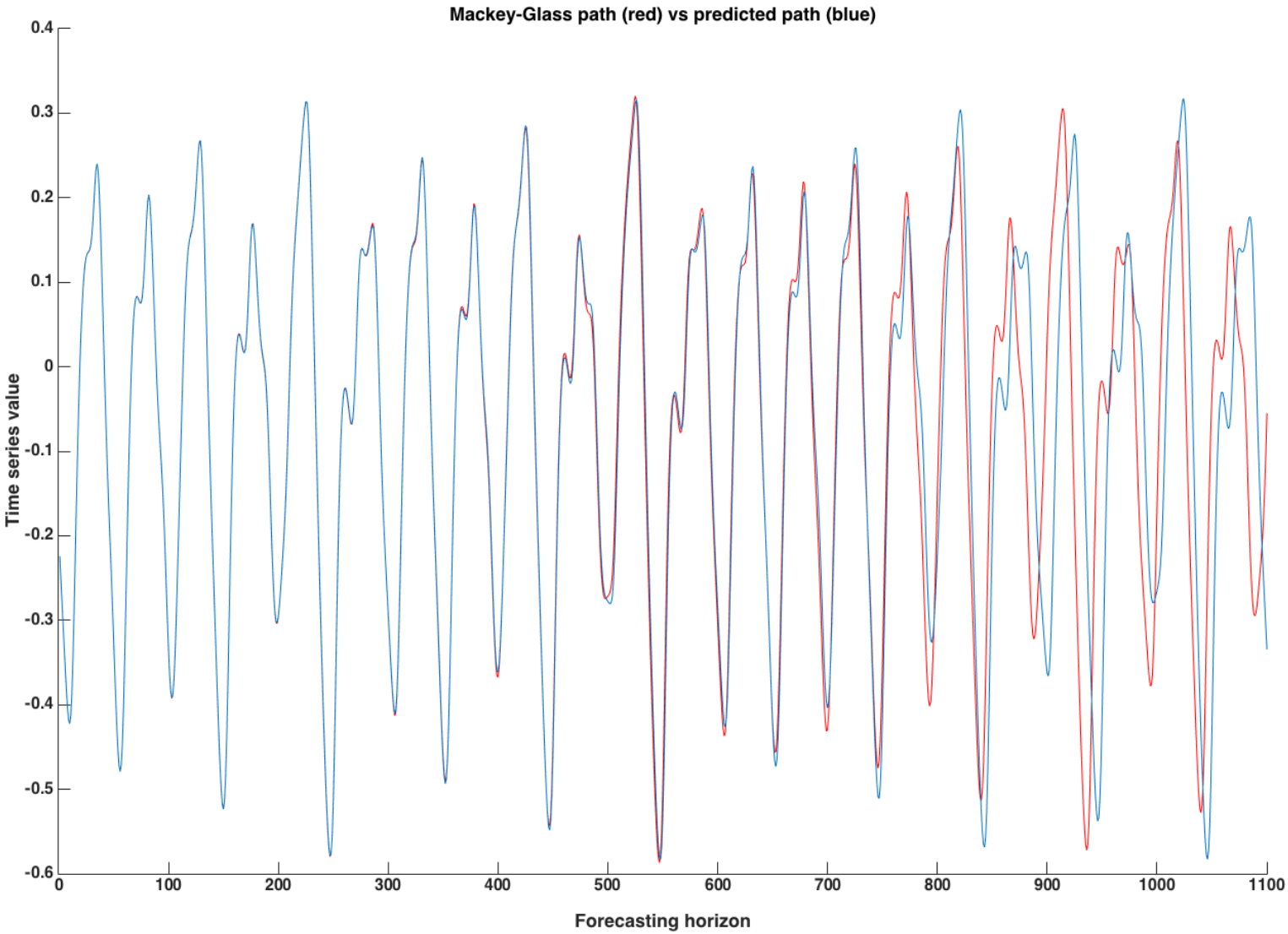
$$\begin{cases} \mathbf{x}_t = \sigma\left(A\mathbf{x}_{t-1} + \mathbf{c}z_t + \mathbf{u}\right), \\ y_t = \mathbf{W}^\top \mathbf{x}_t. \end{cases}$$

Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication

Herbert Jaeger* and Harald Haas

We present a method for learning nonlinear systems, echo state networks (ESNs). ESNs employ artificial recurrent neural networks in a way that has recently been proposed independently as a learning mechanism in biological brains. The learning method is computationally efficient and easy to use. On a benchmark task of predicting a chaotic time series, accuracy is improved by a factor of 2400 over previous techniques. The potential for engineering applications is illustrated by equalizing a communication channel, where the signal error rate is improved by two orders of magnitude.

University of St.Gallen



University of St.Gallen

Learning a chaotic PDE

The Kuramoto-Sivashinsky model for flame propagation

$$y_t = -yy_x - y_{xx} - y_{xxxx} + \mu \cos\left(\frac{2\pi x}{\lambda}\right)$$

- Prediction using an ESN-based learning of this system by Edward Ott's group in [PLH+17, PHG+18]
- It works well up to eight Lyuapunov times

University of St.Gallen

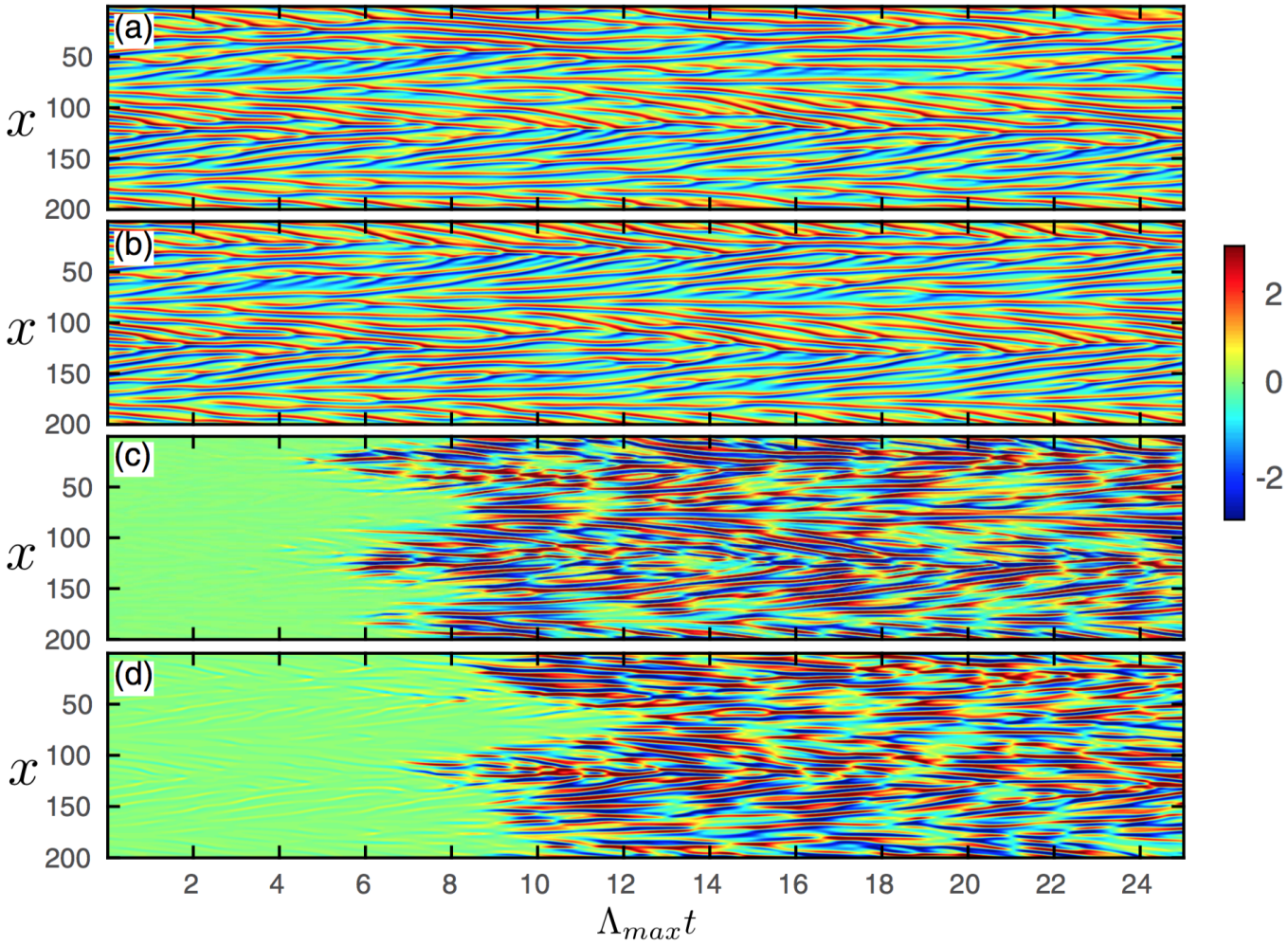
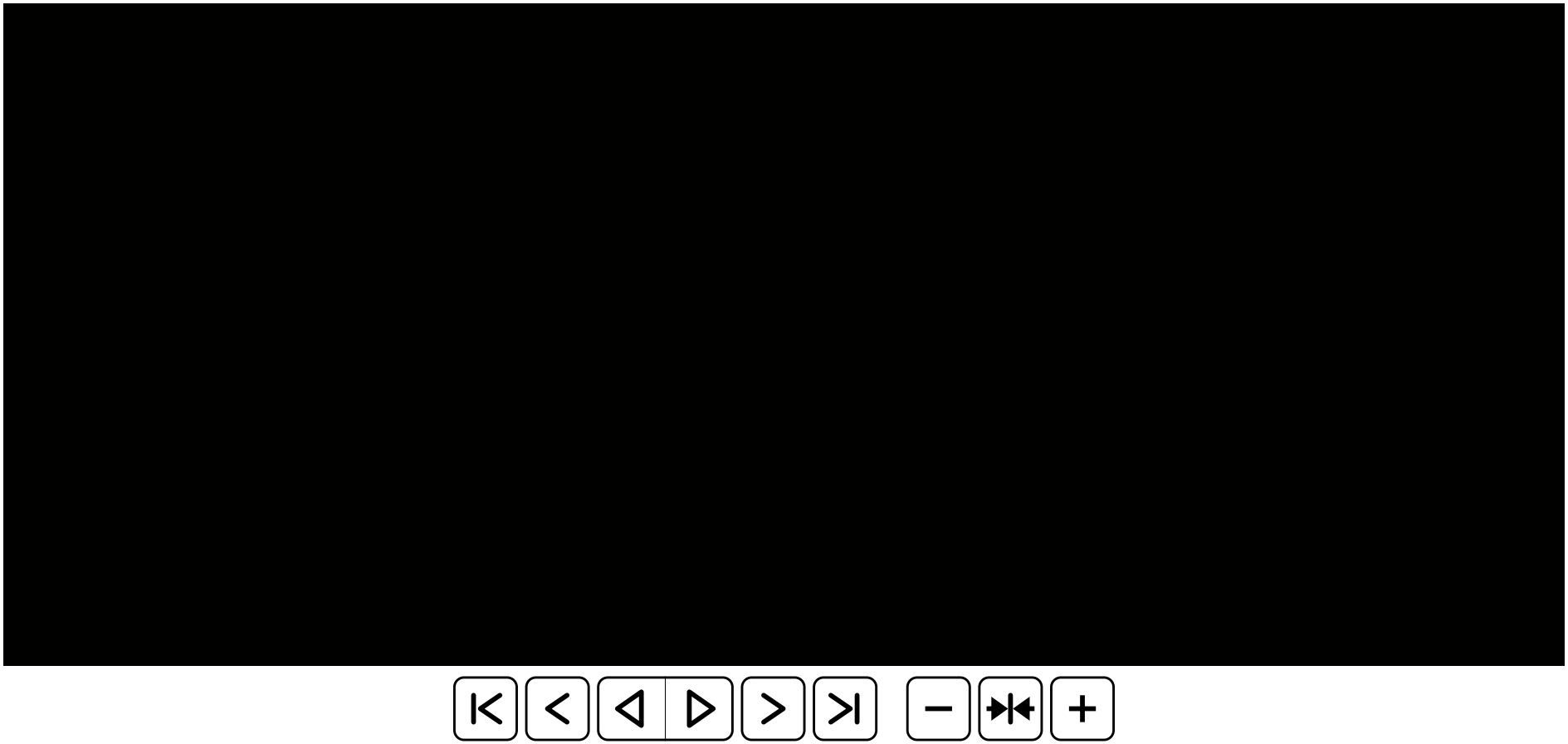


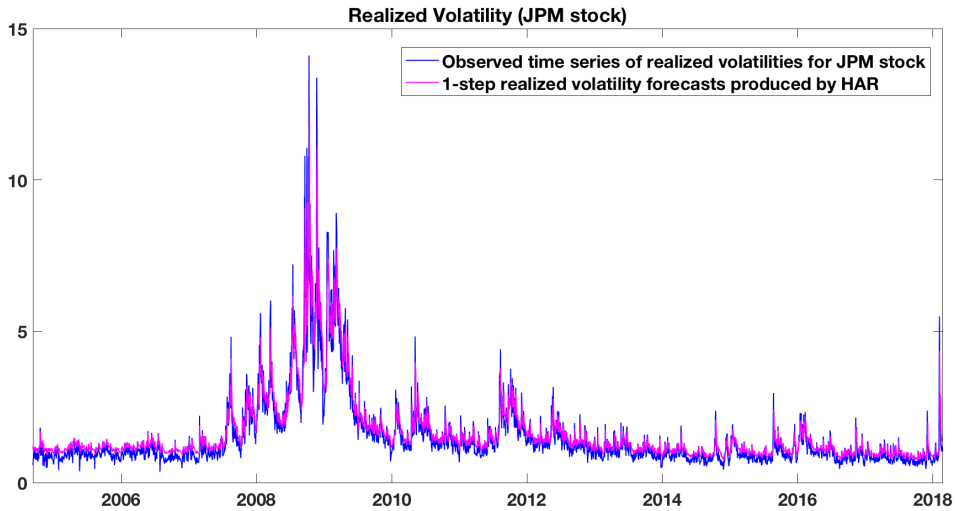
Figure: (a) is the actual solution, (b) is the solution produced by the ESN proxy (c) and (d) are the errors obtained by substraction using two different initializations

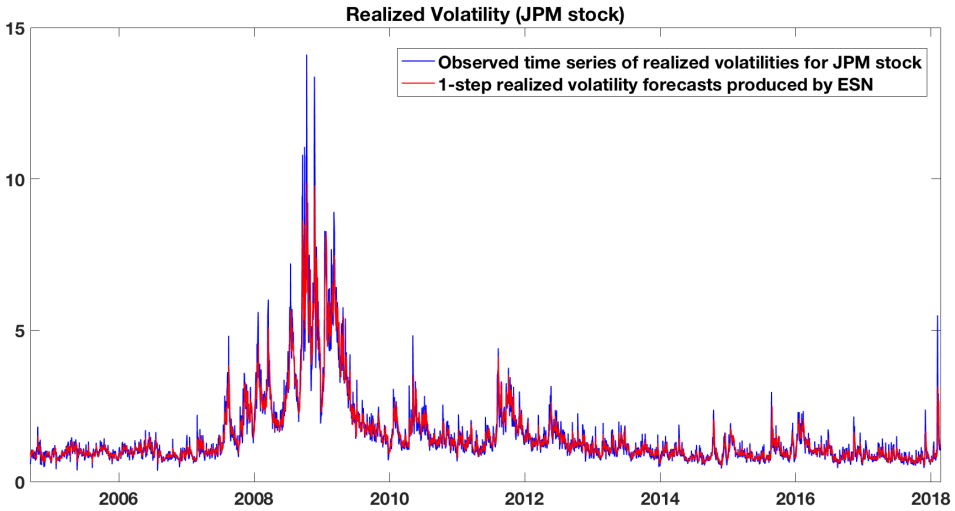
University of St.Gallen



University of St.Gallen

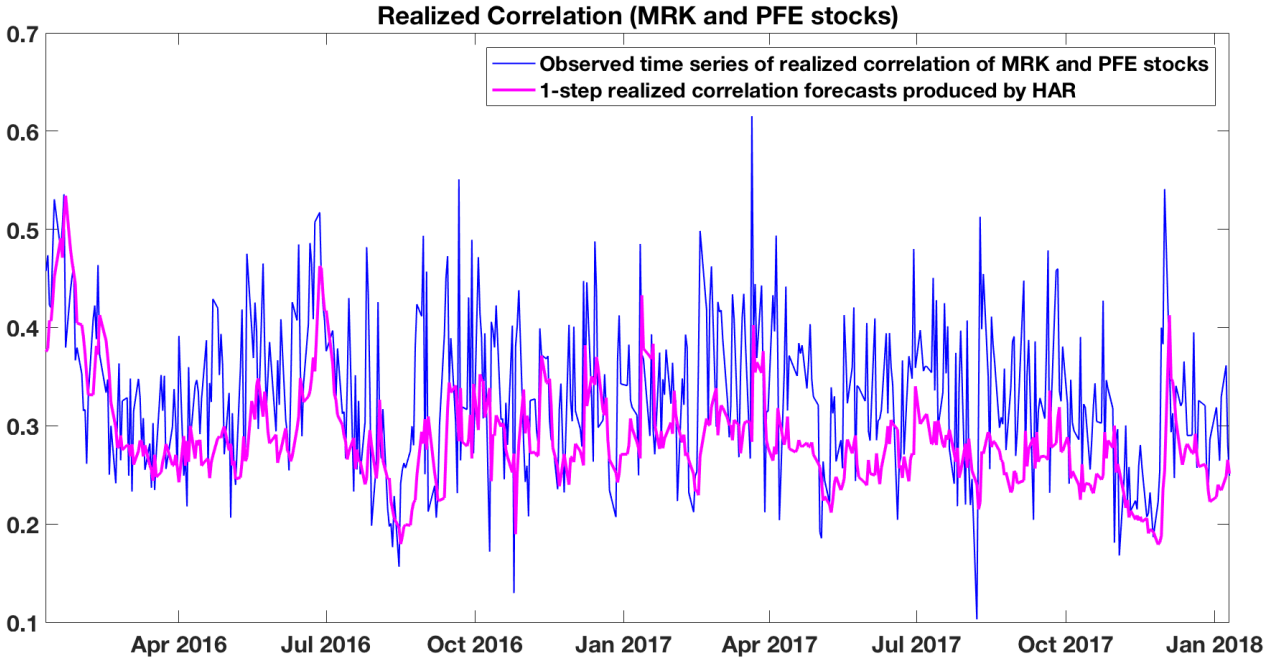
Realized correlation forecasting

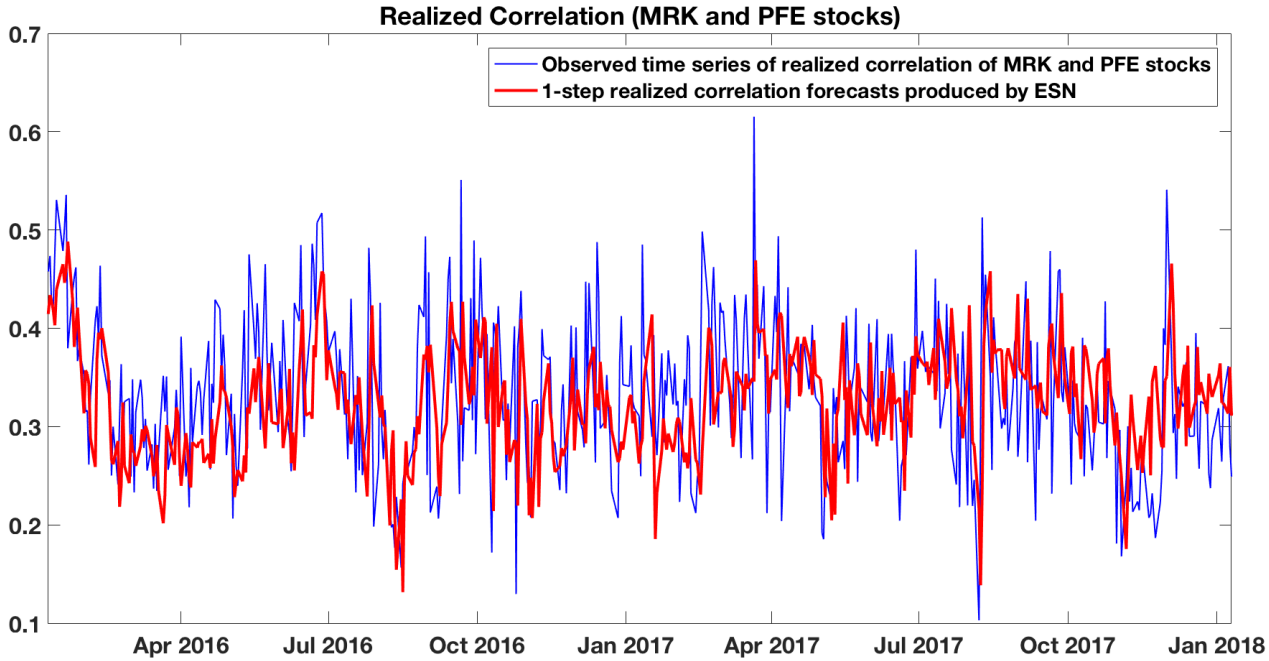




Data cleaning / preprocessing and computation of realized covariances done by Oleksandra Kukharenko; ESN/HAR results are taken from the master thesis of Larissa Zimmermann.

University of St.Gallen





University of St.Gallen

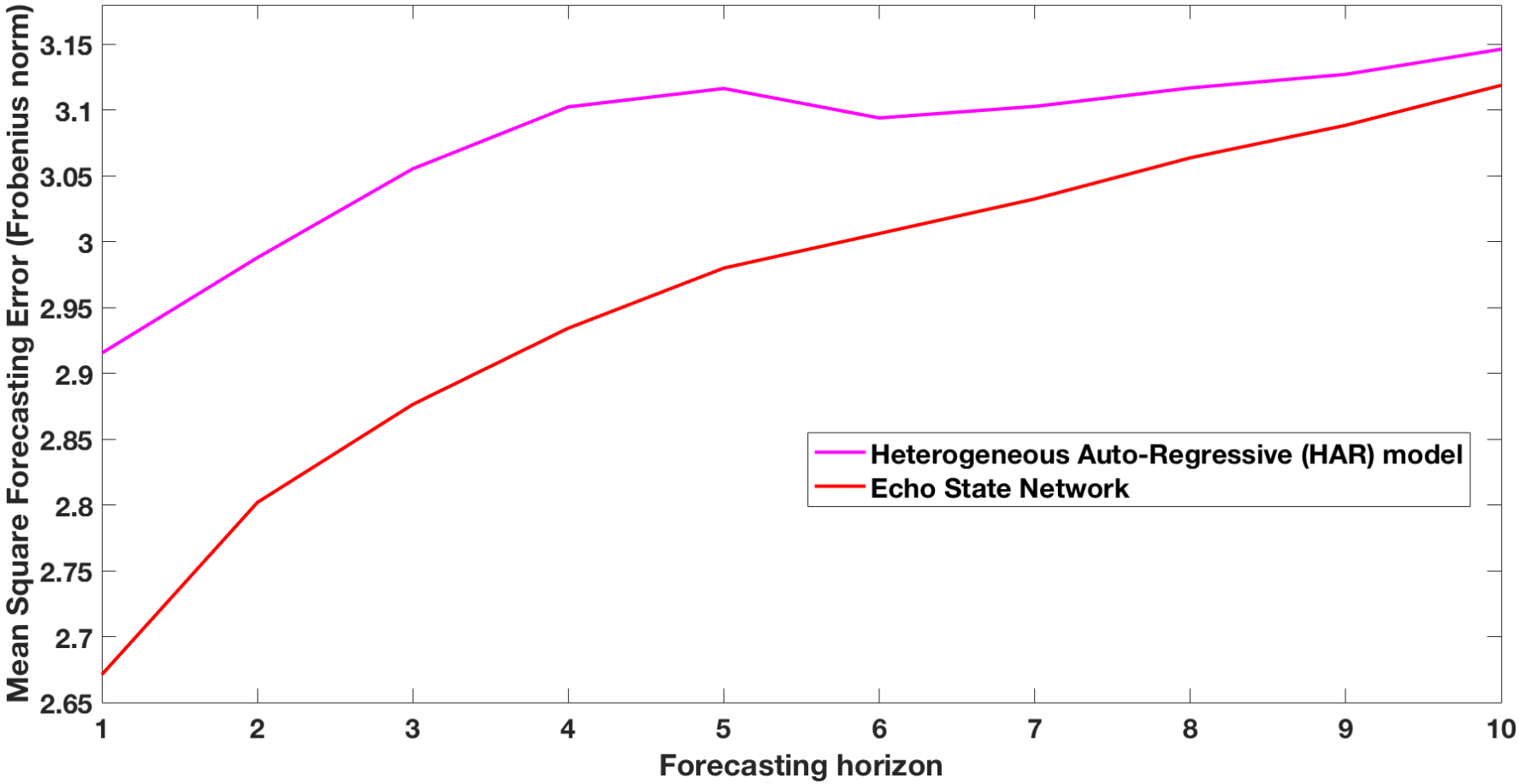


Figure: Multistep forecasting of realized covariance matrices for 4 assets: JPMorgan (JPM), Pfizer (PFE), Merck (MRK), CenturyLink (CTL). Dataset: from 10-Sep-2004 to 21-Feb-2018.

University of St.Gallen

Universality: short literature review

Universality established in different setups with various hypotheses:

- **Continuous time:** available for linear reservoirs with polynomial readouts or for bilinear reservoirs with linear readouts.
  - **Compact time:** corollary of classical results in systems theory by Fliess, Normand-Cyrot, and Sussmann [Fli76, Sus76].
  - **Infinite time:** first formulated by Boyd and Chua [BC85] using the notion of **fading memory**. See also [MS00, MNM02, MNM04, MJS07] for reservoirs coming from the modeling of neural circuits.
- **Discrete and compact time:** when readout is linear required the introduction in [FNC80, DNC84] of the so-called (homogeneous) **state-affine systems (SAS)** (see also [Son79a, Son79b]).
- **Internal approximation approach:** approximating filters via the approximation of the state equations.
  [San91a, San91b, Mat92, Mat93, Per96, SP97].

University of St.Gallen

Our contributions

We extend the previous results to infinite discrete time with stochastic inputs.

1. **Non-homogeneous variant of the state-affine systems (SAS)**: identify sufficient conditions for the associated reservoir computers with linear readouts to be causal, time-invariant, and fading memory.
2. **Universal subset of this class characterized** in the category of fading memory filters with uniformly bounded outputs.
3. **Stochastic setup extension**: version of the universality result that is valid for almost-surely uniformly bounded and measurable inputs with the L^∞ and L^p norms, respectively.
4. **Echo state networks are universal**: this is the dynamic analog of the classical Cybenko and Hornik *et al* theorems in the static setup.

University of St.Gallen

Filters and functionals

Filters $U : (D_n)^{\mathbb{Z}} \longrightarrow \mathbb{R}^{\mathbb{Z}}$ and functionals $H : (D_n)^{\mathbb{Z}} \longrightarrow \mathbb{R}$:

- **Causal filter**: for any two elements $\mathbf{z}, \mathbf{w} \in (D_n)^{\mathbb{Z}}$ that satisfy that $\mathbf{z}_\tau = \mathbf{w}_\tau$ for all $\tau \leq t$, for any given $t \in \mathbb{Z}$, we have that $U(\mathbf{z})_t = U(\mathbf{w})_t$.
- **Time-invariant filter**: when U commutes with the time delay operator U_τ defined by $(U_\tau \mathbf{z})_t := \mathbf{z}_{t-\tau}$, that is, $U_\tau \circ U = U \circ U_\tau$.
- Bijection between causal time-invariant filters and functionals on $(D_n)^{\mathbb{Z}_-}$:

$$\begin{aligned} U &\longrightarrow H_U(\mathbf{z}) := U(\mathbf{z}^e)_0 \\ H &\longrightarrow U_H(\mathbf{z})_t := H((\mathbb{P}_{\mathbb{Z}_-} \circ U_{-t})(\mathbf{z})), \end{aligned}$$

  where U_{-t} is the $(-t)$-time delay operator and $\mathbb{P}_{\mathbb{Z}_-} : (D_n)^{\mathbb{Z}} \longrightarrow (D_n)^{\mathbb{Z}_-}$ is the natural projection. It is easy to verify that:

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

- Let $H_1, H_2 : (D_n)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$, $\lambda \in \mathbb{R}$, then $U_{H_1 + \lambda H_2}(\mathbf{z}) = U_{H_1}(\mathbf{z}) + \lambda U_{H_2}(\mathbf{z})$, for any $\mathbf{z} \in (\mathbb{D}_n)^{\mathbb{Z}}$


University of St.Gallen

Reservoir filters

The reservoir system

$$\begin{cases} \mathbf{x}_t = F(\mathbf{x}_{t-1}, \mathbf{z}_t), & (3) \\ \mathbf{y}_t = h(\mathbf{x}_t), & (4) \end{cases}$$

determines a filter when the following existence and uniqueness property holds (**echo state property** [Jae10, YJK12]): for each $\mathbf{z} \in (D_n)^{\mathbb{Z}}$ there exists a unique $\mathbf{x} \in (D_N)^{\mathbb{Z}}$ such that for each $t \in \mathbb{Z}$, the relation (3) holds.

- The **state filter** $U^F : (D_n)^{\mathbb{Z}} \longrightarrow (D_N)^{\mathbb{Z}}$ is determined by $U^F(\mathbf{z})_t := \mathbf{x}_t \in D_N$
- The **reservoir filter** $U^F_h : (D_n)^{\mathbb{Z}} \longrightarrow \mathbb{R}^{\mathbb{Z}}$ is determined by the entire reservoir system, that is, $U^F_h(\mathbf{z})_t := h\left(U^F(\mathbf{z})_t.\right) = y_t$.

The filters U^F and U^F_h are causal by construction and are necessarily time-invariant [GO18]. We can hence associate to U^F_h a **reservoir functional** $H^F_h : (D_n)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ determined by $H^F_h := H_{U^F_h}$.

University of St.Gallen

Weighted norms, uniformly bounded sequences, compactness

- The **weighted norm** $\|\cdot\|_w$ on $(\mathbb{R}^n)^{\mathbb{Z}_-}$ associated to the **weighting sequence** $w : \mathbb{N} \longrightarrow (0,1]$ as the map:

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

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^n$. The space

$$\ell^\infty_w(\mathbb{R}^n) := \left\{ \mathbf{z} \in (\mathbb{R}^n)^{\mathbb{Z}_-} \mid \|\mathbf{z}\|_w < \infty \right\}, \tag{5}$$

endowed with weighted norm $\|\cdot\|_w$ forms a Banach space [GO18].

- **Two important lemmas**: Let $M > 0$ and let $K_M := \left\{ \mathbf{z} \in (\mathbb{R}^n)^{\mathbb{Z}_-} \mid \|\mathbf{z}_t\| \leq M \text{ for all } t \in \mathbb{Z}_- \right\} = \overline{B_n(\mathbf{0}, M)}^{\mathbb{Z}_-}$. For any weighting sequence w and $\mathbf{z} \in K_M$, we have that K_M is a **compact topological space** when endowed with the relative topology inherited from the norm topology in the Banach space $(\ell^\infty_w(\mathbb{R}^n), \|\cdot\|_w)$.

University of St.Gallen

The fading memory property

- We want filters for which the inputs in the far past do not count.
- We encode the **fading memory property (FMP)** as a continuity property: the causal and time-invariant filter $U : (D_n)^{\mathbb{Z}} \longrightarrow (\mathbb{R})^{\mathbb{Z}}$ has the FMP whenever there exists a weighting sequence $w : \mathbb{N} \longrightarrow (0, 1]$ such that the map $H_U : ((D_n)^{\mathbb{Z}_-}, \|\cdot\|_w) \longrightarrow \mathbb{R}$ is continuous. This means that for any $\mathbf{z} \in (D_n)^{\mathbb{Z}_-}$ and any $\epsilon > 0$, there exists a $\delta(\epsilon) > 0$ such that for any $\mathbf{s} \in (D_n)^{\mathbb{Z}_-}$ that satisfies that

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

  If w is s.t. $w_t = \lambda^t$, for some $\lambda \in (0, 1)$ and all $t \in \mathbb{N}$, then U is said to have the λ-**exponential fading memory property**.
- **FMP does not depend on the weighting sequence:** it can be shown [GO18] that in the case of uniformly bounded input sequences, if a filter has the FMP with respect to a given weighting sequence, it necessarily has the same property with respect to any other weighting sequence.

University of St.Gallen

Universality results in the deterministic setup

Goal: identify families of reservoir filters that are able to uniformly approximate any time-invariant, causal, and fading memory filter with deterministic inputs with any desired degree of accuracy. Such families of reservoir computers are said to be **universal**.

Tools: The Stone-Weierstrass theorem for polynomial subalgebras of real-valued functions defined on compact metric spaces.

Approach: One needs to prove that filters form polynomial algebras. If $D_n \subset \mathbb{R}^n$ and $H_{U_1}, H_{U_2} : (D_n)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ are the functionals associated to the causal and time-invariant filters $U_1, U_2 : (D_n)^{\mathbb{Z}} \longrightarrow \mathbb{R}^{\mathbb{Z}}$, one defines $H_{U_1} \cdot H_{U_2} : (D_n)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ and $H_{U_1} + \lambda H_{U_2} : (D_n)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$, $\lambda \in \mathbb{R}$, as

$$\left(H_{U_1} \cdot H_{U_2}\right)(\mathbf{z}) := H_{U_1}(\mathbf{z}) \cdot H_{U_2}(\mathbf{z}), \quad \left(H_{U_1} + \lambda H_{U_2}\right)(\mathbf{z}) := H_{U_1}(\mathbf{z}) + \lambda H_{U_2}(\mathbf{z}), \tag{6}$$

Theorem

Let

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

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

$$\|H - H^F_h\|_\infty := \sup_{\mathbf{z} \in K_M} |H(\mathbf{z}) - H^F_h(\mathbf{z})| < \epsilon,$$

with $h := p(h_{i_1}, \ldots, h_{i_r})$ and $F := (F_{i_1}, \ldots, F_{i_r})$.

The reservoir systems family is universal

Corollary: The set of all reservoir filters with uniformly bounded inputs in K_M and that have the FMP with respect to a given weighted norm $\|\cdot\|_w$

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

is universal, that is, it is dense in the set $(C^0(K_M), \|\cdot\|_w)$ of real-valued continuous functions on $(K_M, \|\cdot\|_w)$.

Consequence of:

$$H_{h_1}^{F_1} \cdot H_{h_2}^{F_2} = H_h^F, \quad \text{with} \quad h := h_1 \cdot h_2 \in C^\infty(D_{N_1} \times D_{N_2}), \tag{7}$$

$$H_{h_1}^{F_1} + \lambda H_{h_2}^{F_2} = H_{h'}^F, \quad \text{with} \quad h' := h_1 + \lambda h_2 \in C^\infty(D_{N_1} \times D_{N_2}), \tag{8}$$

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

$$F(((\mathbf{x}_1)_t, (\mathbf{x}_2)_t), \mathbf{z}_t) := (F_1((\mathbf{x}_1)_t, \mathbf{z}_t), F_2((\mathbf{x}_2)_t, \mathbf{z}_t)) \tag{9}$$

University of St.Gallen

Linear reservoirs with polynomial readouts are universal

Linear reservoir computer:

$$\begin{cases} \mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{c}\mathbf{z}_t, \quad A \in \mathbb{M}_N, \mathbf{c} \in \mathbb{M}_{N,n}, & (10) \\ y_t = h(\mathbf{x}_t), \quad h \in \mathbb{R}[\mathbf{x}]. & (11) \end{cases}$$

Corollary

The set $\mathcal{L}_\epsilon$ formed by all the linear reservoir systems as in (10)-(11) with matrices $A \in \mathbb{M}_N$ such that $\sigma_{\max}(A) < 1 - \epsilon$ is made of λ_ρ-exponential fading memory reservoir functionals, with $\lambda_\rho := (1-\epsilon)^\rho$, for any $\rho \in (0,1)$. This family is dense in $(C^0(K_M), \|\cdot\|_{w^\rho})$.
The same universality result can be stated for two smaller subfamilies of $\mathcal{L}_\epsilon$ generated by diagonal and nilpotent matrices.

University of St.Gallen

State-affine systems (SAS)

Take two polynomials $p(z) \in \mathbb{M}_{N,N}[z]$ and $q(z) \in \mathbb{M}_{N,1}[z]$ on the variable z with matrix coefficients, that is

$$\begin{aligned} p(z) &:= A_0 + zA_1 + z^2 A_2 + \cdots + z^{n_1} A_{n_1}, \\ q(z) &:= B_0 + zB_1 + z^2 B_2 + \cdots + z^{n_2} B_{n_2} \end{aligned}$$

The **non-homogeneous state-affine system (SAS)** associated to p, q and $\mathbf{W}$ is the reservoir system determined by the state-space transformation:

$$\begin{cases} \mathbf{x}_t = p(z_t)\mathbf{x}_{t-1} + q(z_t), & (12) \\ y_t = \mathbf{W}^\top \mathbf{x}_t. & (13) \end{cases}$$

University of St.Gallen

Integrability of SAS

Proposition

Consider a non-homogeneous SAS defined on $I^{\mathbb{Z}}$, $I := [-1, 1]$. *If* $\max_{z \in I} \|p(z)\|_2 < 1$ *then:*

- *The system has a unique causal and time-invariant solution:*

$$\begin{cases} \mathbf{x}_t = \sum_{j=0}^{\infty} \left(\prod_{k=0}^{j-1} p(z_{t-k}) \right) q(z_{t-j}), & (14) \\ y_t = \mathbf{W}^\top \mathbf{x}_t. & (15) \end{cases}$$

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

- $U^{p,q}_{\mathbf{W}}$ *has the fading memory property.*

SAS form a polynomial algebra

Proposition

$H^{p_1,q_1}_{\mathbf{W}_1}, H^{p_2,q_2}_{\mathbf{W}_2} : D^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ two SAS reservoir functionals. Then:

(i) *Closedness under linear combinations:*

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

(ii) *Closedness under products:*

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

where $p(z)$ the polynomial with matrix coefficients:

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

Theorem (Universality of SAS reservoir computers)

Let

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

and let $\mathcal{S}_\epsilon$ *be the family of functionals* $H^{p,q}_{\mathbf{W}} : I^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ *induced by the state-affine systems in* (12)-(13) *that satisfy that* $M_p := \max_{z\in I} \|p(z)\|_2 < 1-\epsilon$ *and* $M_q := \max_{z\in I} \|q(z)\|_2 < 1-\epsilon$. *The subfamily* $\mathcal{S}_\epsilon$ *is dense in* $(C^0(I^{\mathbb{Z}_-}), \|\cdot\|_{w^\rho})$.
Equivalently, for any fading memory filter H *and any* $\epsilon > 0$, *there exist a natural number* $N \in \mathbb{N}$, *polynomials* $p(z) \in \mathbb{M}_N[z], q(z) \in \mathbb{M}_{N,1}[z]$ *with* $M_p, M_q < 1-\epsilon$, *and a vector* $\mathbf{W} \in \mathbb{R}^N$ *such that*

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

The same universality result can be stated for the smaller subfamily formed by SAS reservoir systems determined by nilpotent polynomials.

Stochastic inputs and outputs

- **Inputs and outputs**: almost surely bounded time series or discrete-time stochastic processes, that is, elements in the space

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

with

$$\|\mathbf{z}\|_{L^\infty}:=\operatorname*{ess\,sup}_{\omega\in\Omega}\|\mathbf{z}(\omega)\|_\infty=\operatorname*{ess\,sup}_{\omega\in\Omega}\left\{\sup_{t\in\mathbb{Z}}\{\|\mathbf{z}_t(\omega)\|\}\right\}. \tag{16}$$

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

- **Weighted norm:** consider w a weighting sequence and let $\|\cdot\|_w$ be the associated weighted norm in $(\mathbb{R}^n)^{\mathbb{Z}_-}$. We work with $L^\infty_w\left(\Omega,(\mathbb{R}^n)^{\mathbb{Z}_-}\right)$, the space of processes $\mathbf{z}:\mathbb{Z}_-\times\Omega$ with finite $\|\cdot\|_{L^\infty_w}$ norm defined as:

$$\|\mathbf{z}\|_{L^\infty_w}:=\operatorname*{ess\,sup}_{\omega\in\Omega}\|\mathbf{z}(\omega)\|_w=\operatorname*{ess\,sup}_{\omega\in\Omega}\left\{\sup_{t\in\mathbb{Z}_-}\{\|\mathbf{z}_t(\omega)\|w_{-t}\}\right\}. \tag{17}$$

Again $L^\infty_w\left(\Omega,(\mathbb{R}^n)^{\mathbb{Z}_-}\right)=L^\infty\left(\Omega,\ell^\infty_w(\mathbb{R}^n)\right)$ and $L^\infty_w\left(\Omega,(\mathbb{R}^n)^{\mathbb{Z}_-}\right)$ is a Banach space.

Deterministic filters in a stochastic setup

- Intrinsically **deterministic filters:** almost surely bounded stochastic inputs $\mathbf{z} \in D_n^{L_{\mathbb{Z}}^\infty} \subset L^\infty(\Omega, (\mathbb{R}^n)^{\mathbb{Z}})$ are presented to filters $U : (D_n)^{\mathbb{Z}} \longrightarrow \mathbb{R}^{\mathbb{Z}}$
- The dependence on the probability space of the image $(U(\mathbf{z}))(\omega)$ takes place exclusively through the dependence $\mathbf{z}(\omega)$ in the input
- Causality/time-invariance of filters are defined as in the deterministic case
- As in the deterministic case there is also a correspondence between causal and time-invariant filters and functionals
- Given a weighting sequence w and a time-invariant filter $U : D_n^{L_{\mathbb{Z}-}^\infty} \longrightarrow L^\infty(\Omega, \mathbb{R}^{\mathbb{Z}})$ with stochastic inputs, one says that U has the **fading memory property** w.r.t. w when the associated functional $H_U : \left(D_n^{L_{\mathbb{Z}-}^\infty}, \|\cdot\|_{L_w^\infty}\right) \longrightarrow L^\infty(\Omega, \mathbb{R})$ is continuous
- Almost surely uniformly bounded inputs/outputs: define

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

University of St.Gallen

The transfer theorem

Theorem (Fading memory and the universality properties inherited by deterministic filters with a.s. bounded inputs)

Let $M > 0$ and let K_M and $K_M^{L^\infty}$ be the sets of deterministic and stochastic inputs, respectively. The following properties hold true:

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

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

University of St.Gallen

Universality via internal approximation

Theorem

Let $K_M \subset (\mathbb{R}^n)^{\mathbb{Z}_-}$ and $K_L \subset (\mathbb{R}^N)^{\mathbb{Z}_-}$ be subsets of uniformly bounded sequences, let $F : \overline{B_{\|\cdot\|}(\mathbf{0}, L)} \times \overline{B_{\|\cdot\|}(\mathbf{0}, M)} \longrightarrow \overline{B_{\|\cdot\|}(\mathbf{0}, L)}$ be a continuous reservoir map.

(i) **Existence of solutions:** *for each $\mathbf{z} \in K_M$ there exists a $\mathbf{x} \in K_L$ (not necessarily unique) that solves the reservoir equation associated to F, that is,*

$$\mathbf{x}_t = F(\mathbf{x}_{t-1}, \mathbf{z}_t), \quad \text{for all } t \in \mathbb{Z}_-.$$

(ii) **Uniqueness and continuity of solutions (ESP and FMP):** *if F is a contraction, then the reservoir system associated to F has the echo state property. Moreover, this system has a unique associated causal and time-invariant filter $U_F : K_M \longrightarrow K_L$ with the fading memory property.*

(iii) **Internal approximation:** *let $F_1, F_2 : \overline{B_{\|\cdot\|}(\mathbf{0}, L)} \times \overline{B_{\|\cdot\|}(\mathbf{0}, M)} \longrightarrow \overline{B_{\|\cdot\|}(\mathbf{0}, L)}$ be continuous reservoir maps s.t. F_1 is a contraction with $0 < r < 1$ and F_2 has the existence of solutions property. Let $U_{F_1}, U_{F_2} : K_M \longrightarrow K_L$ be the corresponding filters. Then, for any $\epsilon > 0$, we have that*

$$\|F_1 - F_2\|_\infty < \delta(\epsilon) := (1-r)\epsilon \quad \text{implies that} \quad \|\|U_{F_1} - U_{F_2}\|\|_\infty < \epsilon. \tag{18}$$

Theorem (Echo state networks are universal)

Let $U : I_n^{\mathbb{Z}_-} \longrightarrow \left(\mathbb{R}^d\right)^{\mathbb{Z}_-}$ be a causal and time-invariant filter that has the fading memory property. Then, for any $\epsilon > 0$ and any weighting sequence w, there is an echo state network

$$\begin{cases} \mathbf{x}_t = \sigma\left(A\mathbf{x}_{t-1} + C\mathbf{z}_t + \boldsymbol{\zeta}\right), & (19) \\ \mathbf{y}_t = W\mathbf{x}_t. & (20) \end{cases}$$

whose associated generalized filters $U_{\mathrm{ESN}} : I_n^{\mathbb{Z}_-} \longrightarrow \left(\mathbb{R}^d\right)^{\mathbb{Z}_-}$ satisfy that

$$\left\| U - U_{\mathrm{ESN}} \right\|_\infty < \epsilon. \quad (21)$$

In these expressions $C \in \mathbb{M}_{N,n}$ for some $N \in \mathbb{N}$, $\boldsymbol{\zeta} \in \mathbb{R}^N$, $A \in \mathbb{M}_{N,N}$, and $W \in \mathbb{M}_{d,N}$. The function $\sigma : \mathbb{R}^N \longrightarrow [-1,1]^N$ in (19) is constructed by componentwise application of a continuous squashing function $\sigma : \mathbb{R} \longrightarrow [-1,1]$.

When the approximating echo state network (19)-(20) satisfies the echo state property, then it has a unique filter U_{ESN} associated which is necessarily time-invariant. The corresponding reservoir functional $H_{\mathrm{ESN}} : I_n^{\mathbb{Z}_-} \longrightarrow \mathbb{R}^d$ satisfies that

$$\left\| H_U - H_{\mathrm{ESN}} \right\|_\infty < \epsilon. \quad (22)$$

Future work

- Performance bounds. Maurey-Barron-Jones Theorems and the curse of dimensionality.
- Capacity estimates.
- We solved the approximation error problem. What about the estimation error problem?
- Relation to time series analysis.

University of St.Gallen

References I

V. I. Arnold.

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

S. Boyd and L. Chua.

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

G. Cybenko.

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

Henri Dang Van Mien and Dorothée Normand-Cyrot.

Nonlinear state affine identification methods: applications to electrical power plants.
Automatica, 20(2):175–188, mar 1984.

Michel Fliess.

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

Michel Fliess and Dorothée Normand-Cyrot.

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

Lyudmila Grigoryeva and Juan-Pablo Ortega.

Echo state networks are universal.
To appear in Neural Networks, 2018.

University of St.Gallen

References II

Kurt Hornik, Maxwell Stinchcombe, and Halbert White.

Multilayer feedforward networks are universal approximators.
Neural Networks, 2(5):359–366, 1989.

Herbert Jaeger.

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

A. N. Kolmogorov.

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

Michael B. Matthews.

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

Michael B. Matthews.

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

Wolfgang Maass, Prashant Joshi, and Eduardo D. Sontag.

Computational aspects of feedback in neural circuits.
PLoS Computational Biology, 3(1):e165, 2007.

W. Maass, T. Natschläger, and H. Markram.

Real-time computing without stable states: a new framework for neural computation based on perturbations.
Neural Computation, 14:2531–2560, 2002.

University of St.Gallen

References III

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

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

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

Jaideep Pathak, Brian Hunt, Michelle Girvan, Zhixin Lu, and Edward Ott.
Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data: A Reservoir Computing Approach.
Physical Review Letters, 120(2):24102, 2018.

Jaideep Pathak, Zhixin Lu, Brian R. Hunt, Michelle Girvan, and Edward Ott.
Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data.
Chaos, 27(12), 2017.

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

Irwin W. Sandberg.
Approximation theorems for discrete-time systems.
IEEE Transactions on Circuits and Systems, 38(5):564–566, 1991.

University of St.Gallen

References IV

Irwin W Sandberg.

Structure theorems for nonlinear systems.
Multidimensional Systems and Signal Processing, 2:267–286, 1991.

E. Sontag.

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

Eduardo D. Sontag.

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

A.R. Stubberud and P.C. Perryman.

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

David A. Sprecher.

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

David A. Sprecher.

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

David A. Sprecher.

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

University of St.Gallen

References V

Héctor J. Sussmann.

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

Izzet B Yildiz, Herbert Jaeger, and Stefan J Kiebel.

Re-visiting the echo state property.
Neural networks : the official journal of the International Neural Network Society, 35:1–9, nov 2012.

University of St.Gallen