Representing dynamics through random dynamical systems

Josef Teichmann

ETH Zürich

Oslo, November 2019

1. CODE
2. Universality in dynamic machine learning
   - Filters
   - System identification
   - Universality in dynamic ML
   - The fading memory category
   - Universality via Stone-Weierstrass
   - Echo state networks are universal
3. Reservoir computing
   - The concept
   - Signature representation theorem
4. Perspectives and Outlook

Introduction

Goal of this talk is ...

- to present the paradigm of reservoir computing and connect it to rNNs and signature re-presentations.
- to apply random projections to construct *true* reservoirs and prove related generalization results.
- to highlight on the role of randomness in learning procedures and to provide some explainations via signature techniques, random projections and time series techniques.

(joint works with Christa Cuchiero, Lukas Gonon, Lyudmila Grigoryeva, Wahid Khosrawi-Sardroudi, Thomas Krabichler, Martin Larsson, and Juan-Pablo Ortega)

Introduction

Goal of this talk is ...

- to present the paradigm of reservoir computing and connect it to rNNs and signature re-presentations.
- to apply random projections to construct *true* reservoirs and prove related generalization results.
- to highlight on the role of randomness in learning procedures and to provide some explainations via signature techniques, random projections and time series techniques.

(joint works with Christa Cuchiero, Lukas Gonon, Lyudmila Grigoryeva, Wahid Khosrawi-Sardroudi, Thomas Krabichler, Martin Larsson, and Juan-Pablo Ortega)

Introduction

Goal of this talk is ...

- to present the paradigm of reservoir computing and connect it to rNNs and signature re-presentations.
- to apply random projections to construct *true* reservoirs and prove related generalization results.
- to highlight on the role of randomness in learning procedures and to provide some explainations via signature techniques, random projections and time series techniques.

(joint works with Christa Cuchiero, Lukas Gonon, Lyudmila Grigoryeva, Wahid Khosrawi-Sardroudi, Thomas Krabichler, Martin Larsson, and Juan-Pablo Ortega)

CODE

First we stay in our comfort zone :-)

We consider differential equations of the form

$$dY_t = \sum_i V_i(Y_t) du_t^i\,,\ Y_0 = y \in E$$

to define evolutions in state space E depending on local characteristics, initial value $y \in E$ and the control u.

CODEs: control as input

For this talk we fix $y \in E$ and consider

$$u \mapsto W\,\mathrm{Evol}_{s,t}(y)$$

and just train the readout and/or the vector fields.

Does this also correspond to classes of networks? Yes: these are continuous time versions of rNNs, LSTMs, etc.

Used for time series, predictions, etc.

Reservoir Computing (RC)

... We aim to learn an input-output map on a high- or infinite dimensional input state space. Consider the input as well as the output as dynamic, e.g. a time series,

Paradigm of Reservoir computing (Herbert Jäger, Lyudmila, Grigoryeva, Wolfgang Maas, Juan-Pablo Ortega, et al.)

Split the input-output map into a generic part of generalized rNN-type (the *reservoir*), which is *not* trained and a readout part, which is trained.

Often the readout is chosen linear and the reservoir has random features. The reservoir is usually a numerically very tractable dynamical system.

Applications of RC

- Often reservoirs can be realized physically, whence ultrafast evaluations are possible. Only the readout map W has to be trained.
- One can learn dynamic phenomena *without* knowing the specific characteristics.
- It works unreasonably well with generalization tasks.

Applications of RC

- Often reservoirs can be realized physically, whence ultrafast evaluations are possible. Only the readout map W has to be trained.
- One can learn dynamic phenomena *without* knowing the specific characteristics.
- It works unreasonably well with generalization tasks.

Applications of RC

- Often reservoirs can be realized physically, whence ultrafast evaluations are possible. Only the readout map W has to be trained.
- One can learn dynamic phenomena *without* knowing the specific characteristics.
- It works unreasonably well with generalization tasks.

An instance of RC in CODEs

Consider a controlled differential equation

$$dY_t = \sum_{i=1}^{d} V_i(Y_t) du_t^i, \ Y_0 = y \in E$$

for some smooth vector fields $V_i : E \to TE$, $i = 1, \dots, d$ and d independent (Stratonovich) Brownian motions u^i, or finite variation continuous controls, or a rough path. This describes a controlled dynamics on E.

We want to learn the dynamics, i.e. the map

(input control u) $\mapsto$ (solution Y).

Obviously a complicated, non-linear map, ...

We introduce some notation for this purpose:

Definition

Let $V : E \to E$ be a smooth vector field, and let $f : E \to \mathbb{R}$ be a smooth function, then we call

$$Vf(x) = df(x) \bullet V(x)$$

the transport operator associated to V, which maps smooth functions to smooth functions and determines V uniquely.

Theorem

Let Evol *be a smooth evolution operator on a convenient vector space E which satisfies (again the time derivative is taken with respect to the forward variable t) a controlled ordinary differential equation*

$$d\,\mathrm{Evol}_{s,t}(x) = \sum_{i=1}^{d} V_i(\mathrm{Evol}_{s,t}(x))du^i(t)$$

then for any smooth function $f : E \to \mathbb{R}$, and every $x \in E$

$$\begin{aligned} & f\big(\,\mathrm{Evol}_{s,t}(x)\big) = \\ & = \sum_{k=0}^{M} \sum_{i_1,\ldots,u_k=1}^{d} V_{i_1} \cdots V_{i_k} f(x) \int_{s\le t_1\le\cdots\le t_k\le t} du^{i_1}(t_1)\cdots du^{i_k}(t_k) + \\ & + R_M(s,t,f) \end{aligned}$$

with remainder term

$$R_M(s,t,f) = \\ = \sum_{i_0,\dots,u_M=1}^{d} \int_{s\le t_1\le\cdots\le t_{M+1}\le t} V_{i_0}\cdots V_{i_k} f\big(\mathrm{Evol}_{s,t_0}(x)\big)\, du^{i_0}(t_0)\cdots du^{i_k}(t_M)$$

holds true for all times $s \le t$ and every natural number $M \ge 0$.

A lot of work has been done to understand the analysis, algebra and geometry of this expansion (Kua-Tsai Chen, Gerard Ben-Arous, Terry Lyons). It is a starting point of *rough path analysis* (Terry Lyons, Peter Friz, etc).

Definition

Consider the free algebra $\mathbb{A}_d$ of formal series generated by d non-commutative indeterminates $e_1, \ldots, e_d$. A typical element $a \in \mathbb{A}_d$ is written as

$$a = \sum_{k=0}^{\infty} \sum_{i_1,\ldots,i_k=1}^{d} a_{i_1 \ldots i_k} e_{i_1} \cdots e_{i_k} ,$$

sums and products are defined in the natural way. We consider the complete locally convex topology making all projections $a \mapsto a_{i_1 \ldots i_k}$ continuous on $\mathbb{A}_d$, hence a convenient vector space.

Definition

We define on $\mathbb{A}_d$ smooth vector fields

$$a \mapsto ae_i$$

for $i = 1, \ldots, d$.

Theorem

Let u be a smooth control, then the controlled differential equation

$$d\,\mathrm{Sig}_{s,t}(a) = \sum_{i=1}^{d} \mathrm{Sig}_{s,t}(a) e_i du^i(t)\,,\ \mathrm{Sig}_{s,s}(a) = a \tag{1}$$

has a unique smooth evolution operator, called signature of u and denoted by Sig*, given by*

$$\mathrm{Sig}_{s,t}(a) = a \sum_{k=0}^{\infty} \sum_{i_1,\ldots,u_k=1}^{d} \int_{s\le t_1\le\cdots\le t_k\le t} du^{i_1}(t_1)\cdots du^{i_k}(t_k)\; e_{i_1}\cdots e_{i_k}\,. \tag{2}$$

Theorem (Signature is a reservoir)

Let Evol *be a smooth evolution operator on a convenient vector space E which satisfies (again the time derivative is taken with respect to the forward variable t) a controlled ordinary differential equation*

$$d\,\mathrm{Evol}_{s,t}(x) = \sum_{i=1}^{d} V_i(\mathrm{Evol}_{s,t}(x))du^i(t)\,.$$

Then for any smooth (test) function $f : E \to \mathbb{R}$ and for every $M \geq 0$ there is a time-homogenous linear $W = W(V_1, \ldots, V_d, f, M, x)$ from $\mathbb{A}_d^M$ to the real numbers $\mathbb{R}$ such that

$$f\big(\,\mathrm{Evol}_{s,t}(x)\big) = W\big(\pi_M(\mathrm{Sig}_{s,t}(1))\big) + \mathcal{O}\big((t-s)^{M+1}\big)$$

for $s \leq t$.

Signature as reservoir

- This explains that any solution can be represented – up to a linear readout – by universal reservoir, namely signature.
- This is used in many instances of provable machine learning by, e.g., groups in Oxford (Harald Oberhauser, Terry Lyons, etc), and also ...
- ... at JP Morgan, in particular great recent work on 'Nonparametric pricing and hedging of exotic derivatives' by Terry Lyons, Sina Nejad and Imanol Perez Arribas.
- in contrast to reservoir computing: signature is high dimensional (i.e. infinite dimensional) and a precisely defined, non-random object.
- Can we approximate signature by a lower dimensional random object with similar properties?

Signature as reservoir

- This explains that any solution can be represented – up to a linear readout – by universal reservoir, namely signature.
- This is used in many instances of provable machine learning by, e.g., groups in Oxford (Harald Oberhauser, Terry Lyons, etc), and also ...
- ... at JP Morgan, in particular great recent work on 'Nonparametric pricing and hedging of exotic derivatives' by Terry Lyons, Sina Nejad and Imanol Perez Arribas.
- in contrast to reservoir computing: signature is high dimensional (i.e. infinite dimensional) and a precisely defined, non-random object.
- Can we approximate signature by a lower dimensional random object with similar properties?

Signature as reservoir

- This explains that any solution can be represented – up to a linear readout – by universal reservoir, namely signature.
- This is used in many instances of provable machine learning by, e.g., groups in Oxford (Harald Oberhauser, Terry Lyons, etc), and also ...
- ... at JP Morgan, in particular great recent work on 'Nonparametric pricing and hedging of exotic derivatives' by Terry Lyons, Sina Nejad and Imanol Perez Arribas.
- in contrast to reservoir computing: signature is high dimensional (i.e. infinite dimensional) and a precisely defined, non-random object.
- Can we approximate signature by a lower dimensional random object with similar properties?

Signature as reservoir

- This explains that any solution can be represented – up to a linear readout – by universal reservoir, namely signature.
- This is used in many instances of provable machine learning by, e.g., groups in Oxford (Harald Oberhauser, Terry Lyons, etc), and also ...
- ... at JP Morgan, in particular great recent work on 'Nonparametric pricing and hedging of exotic derivatives' by Terry Lyons, Sina Nejad and Imanol Perez Arribas.
- in contrast to reservoir computing: signature is high dimensional (i.e. infinite dimensional) and a precisely defined, non-random object.
- Can we approximate signature by a lower dimensional random object with similar properties?

Signature as reservoir

- This explains that any solution can be represented – up to a linear readout – by universal reservoir, namely signature.
- This is used in many instances of provable machine learning by, e.g., groups in Oxford (Harald Oberhauser, Terry Lyons, etc), and also ...
- ... at JP Morgan, in particular great recent work on 'Nonparametric pricing and hedging of exotic derivatives' by Terry Lyons, Sina Nejad and Imanol Perez Arribas.
- in contrast to reservoir computing: signature is high dimensional (i.e. infinite dimensional) and a precisely defined, non-random object.
- Can we approximate signature by a lower dimensional random object with similar properties?

It is the assertion of the Johnson-Lindenstrauss (JL) Lemma that for every $0 < \epsilon < 1$ an N point set Q in some arbitrary (scalar product) space H, can be embedded into a space $\mathbb{R}^k$, where $k = \frac{24 \log N}{3\epsilon^2 - 2\epsilon^3}$ in an almost isometric manner, i.e. there is a linear map $f : H \to \mathbb{R}^k$ such that

$$(1 - \epsilon)\|v_1 - v_2\|^2 \leq \|f(v_1) - f(v_2)\|^2 \leq (1 + \epsilon)\|v_1 - v_2\|^2$$

for all $v_1, v_2 \in Q$. It is remarkable that f can be chosen randomly from a set of linear projection maps and the choice satisfies the desired requirements with high probability.

The result is due to concentration of measure results in high dimensional spaces and has been discovered in the eighties, for some details see below.

In order to make this program work, we need a definition:

Definition

Let Q be any (finite or infinite) set of elements of norm one in $\mathbb{A}_d^M$. For $v \in \mathbb{A}_d^M$ we define the function

$$\|v\|_Q := \inf \left\{ \sum_j |\lambda_j| \mid \sum_j \lambda_j v_j = v \text{ and } v_j \in Q \right\}.$$

We use the convention $\inf \emptyset = +\infty$ since the function is only finite on $\mathrm{span}(Q)$. Actually the function $\|.\|_Q$ behaves precisely like a norm on the span. Additionally $\|v\|_{Q_1} \geq \|v\|_{Q_2}$ for $Q_1 \subset Q_2$ and $\|v\|_Q \geq \|v\|$ for all sets Q of elements of norm one.

Proposition

Fix $M \geq 1$, $\epsilon > 0$ and consider the free nilpotent algebra $\mathbb{A}_d^M$. Let $Q = -Q$ be any N point set of vectors with norm one, then there is linear map $f : \mathbb{A}_d^M \to \mathbb{R}^k$ (k being the above JL constant with N), such that

$$\left|\langle v_1, v_2 - (f^* \circ f)(v_2)\rangle\right| \leq \epsilon\,,$$

for all $v_1, v_2 \in Q$. In particular

$$\left|\langle v_1, v_2 - (f^* \circ f)(v_2))\rangle\right| \leq \epsilon \|v_1\|_Q \|v_2\|_Q\,,$$

for $v_1, v_2 \in \mathbb{A}_d^M$.

Theorem (Cuchiero, Gonon, Grigoryeva, Ortega, Teichmann (2019))

Let u be a smooth control and f the previously constructed JL map associated to a spanning N point set Q of norm one. We denote by r-Sig *the smooth evolution of*

$$dZ_t = \sum_{i=1}^{d} f(f^*(Z_t)e_i)du^i(t)\,,\ Z_0 = f^*(1)$$

a controlled differential equation on $\mathbb{R}^k$. Then

$$\begin{aligned}
&\langle u, \mathrm{Sig}_{s,t}(1) - f^*(\text{r-Sig}_{s,t}(1))\rangle \\
&\leq \Big(\left|\langle \Gamma_{\mathrm{Sig}_{s,t}(1)}(u), 1 - f^*(f(1))\rangle\right| + \\
&+ C\epsilon \sum_{i=1}^{d} \int_s^t \|\Gamma_{\mathrm{Sig}_{r,t}(1)}(u)\|_Q \|f^*(Y_r)e_i\|_Q \, dr \Big)\,,
\end{aligned}$$

with constant $C = \sup_{s\leq r\leq t,\, i} \left|\frac{du^i(r)}{dr}\right|$, and for each $u \in Q$.

Corollary

Let u be a smooth control and f the previously constructed JL map associated to a spanning N point set Q of norm one. Assume additionally $1 = f^*(f(1))$, then

$$\left\| \mathrm{Sig}_{s,t}(1) - f^*(\text{r-Sig}_{s,t}(1)) \right\| \leq$$

$$\left(\epsilon C \sum_{i=1}^{d} \int_s^t \sup_{\|u\|=1} \left\| \Gamma_{\mathrm{Sig}_{r,t}(1)}(u) \right\|_Q \| f^*(Y_r) e_i \|_Q \, dr \right).$$

Hence f^*(r-Sig) approximates Sig *up to order ϵ and can be used as a proxy for signature.*

r-Sig is a random dynamical system

It is fascinating that we can actually calculate approximately the vector fields which determine the dynamics of r-Sig, i.e.

$$y \mapsto f(f^*(y)e_i)$$

for each $i = 1, \ldots, d$ for $y \in \mathbb{R}^k$.

Theorem

For $M \to \infty$ the linear vector fields

$$y \mapsto f(f^*(y)e_i)$$

for $i = 1, \ldots, d$, are built from matrices on $\mathbb{R}^k$ with asymptotically normally distributed, (almost) independent entries.

Randomness matters

Consider

$$dY_t = \sum_{i=1}^{d} V_i(Y_t) du^i(t)\,,\ Y_0 \in E$$

where we observe *one* trajectory on $[0, T]$ and do not know the characteristics.

Randomized Signature

A random localized signature

- there is a set of hyper-parameters $\theta \in \Theta$, and a dimension M.
- depending on θ choose randomly matrices $A_1, \ldots, A_d$ on $\mathbb{R}^M$ as well as shifts $\beta_1, \ldots, \beta_d$ such that maximal non-integrability holds on a starting point $x \in \mathbb{R}^M$.
- one can tune the hyper-parameters $\theta \in \Theta$ and dimension M such that

$$dX_t = \sum_{i=1}^{d} \sigma(A_i X_t + \beta_i) du^i(t)\,,\; X_0 = x$$

locally (in time, as well as space) approximates CODE Y via a linear readout W up to arbitrary precision. σ is a sigmoid function whose only role is to localize the meaning of signature: outside a certain ball the system is not expressive anymore.

Problems of the presented approach

- Signature as well as randomized signature are regression bases of a polynomial type, whence they also come with the problems of polynomial regressions, e.g. bad generalization properties.
- If the dynamical system Y has stationarity properties, it is clear that a representation through signature is sub-optimal, since it does *not* have inherent stationarity properties. Randomized signature can help, but has to be tuned.
- Signature as well as randomized signature are continuous time objects, which are used for discrete time time series by interpolation. Can we directly construct signature like objects in discrete time, possibly with stationarity properties?

Problems of the presented approach

- Signature as well as randomized signature are regression bases of a polynomial type, whence they also come with the problems of polynomial regressions, e.g. bad generalization properties.
- If the dynamical system Y has stationarity properties, it is clear that a representation through signature is sub-optimal, since it does *not* have inherent stationarity properties. Randomized signature can help, but has to be tuned.
- Signature as well as randomized signature are continuous time objects, which are used for discrete time time series by interpolation. Can we directly construct signature like objects in discrete time, possibly with stationarity properties?

Problems of the presented approach

- Signature as well as randomized signature are regression bases of a polynomial type, whence they also come with the problems of polynomial regressions, e.g. bad generalization properties.
- If the dynamical system Y has stationarity properties, it is clear that a representation through signature is sub-optimal, since it does *not* have inherent stationarity properties. Randomized signature can help, but has to be tuned.
- Signature as well as randomized signature are continuous time objects, which are used for discrete time time series by interpolation. Can we directly construct signature like objects in discrete time, possibly with stationarity properties?

Leaving the comfort zone ...

We enter the world of time series analysis:

- the control u input is a discrete path indexed on $\mathbb{Z}_{\leq 0}$. Its values stay in a compact set $D_d \subset \mathbb{R}^d$.
- the state space E is some bounded subset of a finite dimensional space $\mathbb{R}^N$.
- in order to guarantee solutions of such system stationarity is a necessary requirement: no initial values are given.

Leaving the comfort zone ...

We enter the world of time series analysis:

- the control u input is a discrete path indexed on $\mathbb{Z}_{\leq 0}$. Its values stay in a compact set $D_d \subset \mathbb{R}^d$.
- the state space E is some bounded subset of a finite dimensional space $\mathbb{R}^N$.
- in order to guarantee solutions of such system stationarity is a necessary requirement: no initial values are given.

Leaving the comfort zone ...

We enter the world of time series analysis:

- the control u input is a discrete path indexed on $\mathbb{Z}_{\leq 0}$. Its values stay in a compact set $D_d \subset \mathbb{R}^d$.
- the state space E is some bounded subset of a finite dimensional space $\mathbb{R}^N$.
- in order to guarantee solutions of such system stationarity is a necessary requirement: no initial values are given.

Filters and functionals

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

- **Causal filter**: for any two elements $\mathbf{z}, \mathbf{w} \in (D_d)^{\mathbb{Z}}$ 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 $(D_d)^{\mathbb{Z}_-}$:

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

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

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

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

External system identification approach

One models directly the input/output system:

- **Linear systems (LTIs)**: The LTI system $U_{\mathbf{h}} : \ell^{\infty}_{-}(\mathbb{R}) \longrightarrow \ell^{\infty}_{-}(\mathbb{R})$ determined by $\mathbf{h} \in \ell^{1}_{-}(\mathbb{R})$ is defined as

$$U_{\mathbf{h}}(\mathbf{z})_t = \sum_{j \in \mathbb{Z}_-} h_j z_{t+j} =: (\mathbf{h} * \mathbf{z})(t),$$

  where $\mathbf{h} * \mathbf{z}$ is the **convolution product** of $\mathbf{h}$ and $\mathbf{z}$. Estimation via impulse/response analysis. Not every LTI has a convolution representation (Kantorovich, Akilov (1982)).

- **Nonlinear systems and Volterra series**: polynomial expansions of the form

$$U(\mathbf{z})_t = \sum_{j=1}^{\infty} \sum_{m_1=-\infty}^{0} \cdots \sum_{m_j=-\infty}^{0} g_j(m_1, \ldots, m_j) z_{m_1+t} \cdots z_{m_j+t}, \quad t \in \mathbb{Z}_-.$$

  lack of parsimony, poor generalization properties.

Internal system identification approach

Internal variables $\mathbf{x}_t$ are introduced. Filter is modeled as a state-space system

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

- Classical problem in **systems** and **control theory** with profound influence in the development of modern engineering sciences: the **realization problem**. Kalman, Sussmann, Sontag for deterministic. Akaike, Ruckebusch, Lindquist, Picci for stochastic.
- Internal variables may be just auxiliary or encode unobservable factors (stochastic volatility models and **filtering theory**)
- Can be seen as a **non-autonomous dynamical system** or a **recurrent neural network** (rNN)
- State-space systems are **recursively computable**.

The Echo State Property

Internal Representation $\Longrightarrow$ External Representation

State-space systems determine a filter when the following existence and uniqueness property holds (**echo state property (ESP)**): for each $\mathbf{z} \in (D_d)^{\mathbb{Z}}$ there exists a unique $\mathbf{x} \in (D_N)^{\mathbb{Z}}$ such that for each $t \in \mathbb{Z}$, the relation above holds.

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

The filters U^F and U_h^F are causal and time-invariant. H_h^F is the **reservoir functional** determined by $H_h^F := H_{U_h^F}$.

Finite and infinite dimensional state spaces

External Representation $\Longrightarrow$ Internal Representation

Not every time-invariant causal filter admits a finite dimensional realization:

- Short term memory processes ($\gamma(h) \sim a^h$ with $|a| < 1$) **do** (ARIMA)
- Long memory processes ($\rho(h) \sim h^{2d-1}$, $0 < d < 1/2$ (ARFIMA), Hosking (1981)) **don't** (Chan, Palma (1998))
- In the reservoir computing context RKHS state spaces: Hermans, Schrauwen (2012), Hamzi (2019), Tiňo (2019).

We formulate universality results for causal time-invariant filters in discrete time with semi-infinite input.

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. **Echo state networks are universal**: this is the dynamic analog of the classical Cybenko and Hornik *et al* theorems in the static setup.

The fading memory property

- Modeling property introduced by Volterra and von Neumann.
- We want filters for which the inputs in the far past count less than recent ones.
- The **weighted norm** $\|\cdot\|_w$ on $(\mathbb{R}^d)^{\mathbb{Z}_-}$ associated to the **weighting sequence** $w : \mathbb{N} \longrightarrow (0,1]$ as the map:

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

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

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

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

The fading memory property

- We encode the **fading memory property (FMP)** as a continuity property: the causal and time-invariant filter $U : (D_d)^{\mathbb{Z}} \longrightarrow (\mathbb{R})^{\mathbb{Z}}$ has the FMP whenever there exists a weighting sequence $w : \mathbb{N} \longrightarrow (0, 1]$ such that the map $H_U : ((D_d)^{\mathbb{Z}_-}, \|\cdot\|_w) \longrightarrow \mathbb{R}$ is continuous. This means that for any $\mathbf{z} \in (D_d)^{\mathbb{Z}_-}$ and any $\epsilon > 0$, there exists a $\delta(\epsilon) > 0$ such that for any $\mathbf{s} \in (D_d)^{\mathbb{Z}_-}$ that satisfies that

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

- **FMP does not depend on the weighting sequence:** 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.

The fading memory property

- It prevents pathological phenomena: LTIs admit a convolution representation iff they are FMP (Boyd [1985])
- **Important lemma**: Let $M > 0$ and let

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

K_M is a **compact topological space** when endowed with the relative topology inherited from $(\ell_w^\infty(\mathbb{R}^d), \|\cdot\|_w)$.

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 $D_d \subset \mathbb{R}^n$ and $H_{U_1}, H_{U_2} : (D_d)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ are the functionals associated to the causal and time-invariant filters $U_1, U_2 : (D_d)^{\mathbb{Z}} \longrightarrow \mathbb{R}^{\mathbb{Z}}$, one defines $H_{U_1} \cdot H_{U_2} : (D_d)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ and $H_{U_1} + \lambda H_{U_2} : (D_d)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$, $\lambda \in \mathbb{R}$, as

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

The reservoir systems family is universal

Theorem: 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 $\|\cdot\|_w$

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

is universal, that is, it is dense in the set $(C^0(K_M), \|\cdot\|_w)$.

Consequence of:

$$\begin{aligned} H_{h_1}^{F_1} \cdot H_{h_2}^{F_2} &= H_h^F, \quad \text{with} \quad h := h_1 \cdot h_2 \in C^\infty(D_{N_1} \times D_{N_2}), \\ H_{h_1}^{F_1} + \lambda H_{h_2}^{F_2} &= H_{h'}^F, \quad \text{with} \quad h' := h_1 + \lambda h_2 \in C^\infty(D_{N_1} \times D_{N_2}), \end{aligned}$$

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

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

Linear reservoirs with polynomial readouts are universal

Linear reservoir computer:

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

Corollary

The set $\mathcal{L}_\epsilon$ formed by all the linear reservoir systems 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.

Universality with linear readouts: 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:

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

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:*

$$\begin{cases} \mathbf{x}_t = \sum_{j=0}^{\infty} \left(\prod_{k=0}^{j-1} p(z_{t-k}) \right) q(z_{t-j}), & (3) \\ y_t = \mathbf{W}^{\top} \mathbf{x}_t. & (4) \end{cases}$$

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

Universality via internal approximation

Theorem

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*

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

Theorem (Echo state networks are universal)

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

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

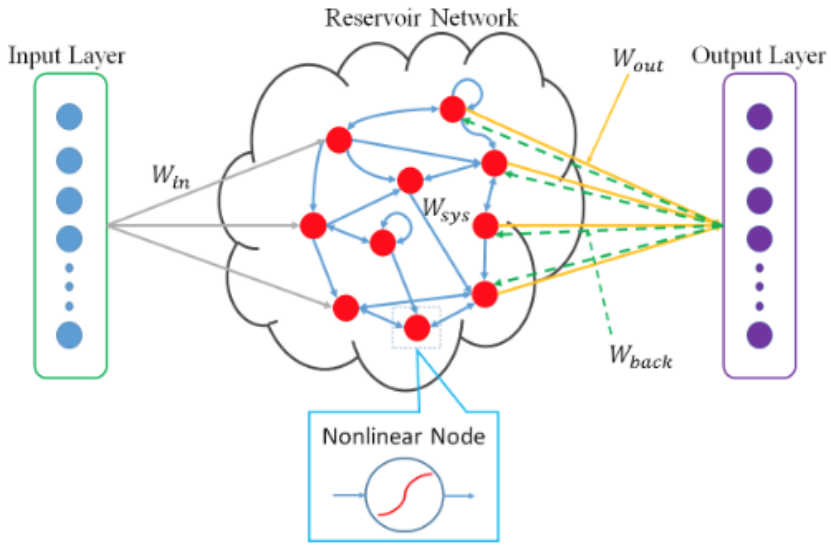
whose associated generalized filters $U_{\mathrm{ESN}} : I_d^{\mathbb{Z}_-} \longrightarrow \left(\mathbb{R}^d\right)^{\mathbb{Z}_-}$ satisfy that

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

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

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

Reservoir computing



- State-space system in which the state function is **chosen quasi-randomly** and remains unchanged during the training stage.
- Readout is **preferably linear** so that data intensive applications become tractable.
- Availability of realizations with dedicated hardware.

Is this reasonable?

We start with the **linearity** requirement and show that any analytic filter admits state-space representation with linear readouts. This is carried out by using a **discrete-time signature process** on the free algebra of infinite series in d indeterminates $e_1 \dots, e_d$

$$\mathbb{A}_d = \Big\{ \sum_{k=0}^{\infty} \sum_{i_1,\dots,i_k=1}^{d} a_{i_1 \cdots i_k} e_{i_1} \cdots e_{i_k} \Big\}$$

This is an infinite dimensional, complete locally convex algebra with component-wise convergence. Denote by

$$\mathbb{A}_d^M = \pi^M(\mathbb{A}_d)$$

the canonical projection on series up to degree M.

The discrete-time signature process

Proposition

Let $0 < \lambda < 1$ and let $r > 0$ be such that $r < (1 - \sqrt{\lambda})/\lambda$. Consider the state-space system on $\mathbb{A}_d$ with inputs in $V_r := \overline{B_{\|\cdot\|_1}(\mathbf{0}, r)} \subset \ell_-^\infty(\mathbb{R}^d)$ given by

$$\mathbf{x}_t = \lambda \mathbf{x}_{t-1} + \lambda \mathbf{x}_{t-1} z_t + z_t.$$

Let $L := 1/\lambda(1 - \lambda r)$ and $V_L := \overline{B_{\|\cdot\|_\infty}(\mathbf{0}, L)} \subset \ell_-^\infty(\mathbb{A}_d)$. This system has the (V_r, V_L)-ESP and determines a unique state filter $U^S : (V_r, \|\cdot\|_\infty) \longrightarrow (V_L, \|\cdot\|_\infty)$ given by

$$U^S(\mathbf{z})_t := \sum_{j=1}^{\infty} \sum_{m_1=-\infty}^{0} \sum_{m_2=-\infty}^{m_1-1} \cdots \sum_{m_j=-\infty}^{m_{j-1}-1} \lambda^{|m_j|} z_{t+m_j} \cdots z_{t+m_1}. \tag{5}$$

Signature representation theorem

Theorem (Cuchiero, Gonon, Grigoryeva, Teichmann, JPO (2019))

Any analytic filter $U : (V_r, \|\cdot\|_\infty) \longrightarrow (V_L, \|\cdot\|_\infty)$ *admits a reservoir filter representation through the signature state filter* U^S*, that is, the filter* U *determines a unique densely defined linear operator* $W : \mathbb{A}_d \to \mathbb{R}^N$ *such that*

$$U(z)_t = W(U^S(z)_t).$$

Let U^S_M *be the finite dimensional truncation of the signature filter at level* M*. Then*

$$\left\| U(\mathbf{z}) - W\pi^M(U^S_M(\mathbf{z})) \right\|_\infty \leq L \left(1 - \frac{\|\mathbf{z}\|_\infty}{r}\right)^{-1} \left(\frac{\|\mathbf{z}\|_\infty}{r}\right)^{M+1}.$$

- The truncated filter is the reservoir filter associated to the linear readout induced by $W \circ \pi^M$ and the reservoir map $F_M : \pi^M(\overline{B_{\|\cdot\|}(0,L)}) \times \overline{B_{\|\cdot\|}(0,r)} \longrightarrow \pi^M(\overline{B_{\|\cdot\|}(0,L)})$ defined by

$$F^M(\overline{\mathbf{x}}, z) := \pi^M F((\pi^M)^*(\overline{\mathbf{x}}), z) = \lambda \overline{\mathbf{x}} + \lambda \pi^M(\overline{\mathbf{x}} z) + z.$$

  Equivalently,

$$U^S_M = U^{F^M} = \pi^M(U^S)$$

- This can be read by saying that the (truncated) Volterra series expansion of any analytic filter coincides with the unique solution of the state-space system in the proposition (respectively, above).
- Tremendous numerical efficiency: already in discrete time and no need to compute infinite sums.

What about randomness?

Two possible answers:

- **Random projections and Johnson-Lindenstrauss (JL) Lemma**: finite dimensional random projections of the signature process are able to approximate analytic filters with arbitrarily small error and high probability.
- Formulation of **PAC-type bounds** on the estimation and approximation errors committed when using randomly chosen elements within a given family of reservoir systems: ESNs, SAS.

Both answers yield **randomly chosen reservoir systems with linear readouts** that approximate analytic filters with arbitrarily small error and high probability. **The reservoir computing paradigm is mathematically reasonable**.

JL reduction of truncated signature representations

Given a Hilbert space H, a N-point set Q in it, and $0 < \epsilon < 1$, there is a linear map $f : H \longrightarrow \mathbb{R}^k$ with $k = \left\lfloor \frac{24 \log N}{3\epsilon^2 - 2\epsilon^3} \right\rfloor$ that approximately preserves the distances of the points in Q:

$$(1-\epsilon)\left\|v_1 - v_2\right\|^2 \leq \left\|f(v_1) - f(v_2)\right\|^2 \leq (1+\epsilon)\left\|v_1 - v_2\right\|^2$$

The projection is randomly drawn from a distribution (standard Gaussian on the entries) on the space of matrices with the right dimensions and the probability that the above inequalities are satisfied is bounded below by $1 - (1/N^2)$

Corollary

Let $W\pi^M U_M^S$ *be the truncated signature representation of the analytic filter* $U : (V_r, \|\cdot\|_\infty) \longrightarrow (V_L, \|\cdot\|_\infty)$. *Then, for any* $0 < \epsilon < 1$ *and sufficiently large* $M \in \mathbb{N}$, *there exists a (random) projection*

$$f : \mathbb{A}_d^M \longrightarrow \mathbb{R}^K, \quad \textit{with} \quad K = \left\lfloor \frac{24 \log(\dim \mathbb{A}_d^M)}{3\epsilon^2 - 2\epsilon^3} \right\rfloor$$

such that the reservoir map $\widehat{F_K} : S_K \times \overline{B_{\|\cdot\|}(0, r)} \longrightarrow S_K$ *with* $S_K := f^* f \left(\pi^M(\overline{B_{\|\cdot\|}(0, L)}) \right)$ *and given by*

$$\widehat{F_K}(\widehat{\mathbf{x}}, \mathbf{z}) = f^* f \left(\lambda \widehat{\mathbf{x}} + \lambda \sum_{i=1}^{d} z^i \pi^M(\widehat{\mathbf{x}} \mathbf{e}_i) + z^i e_i \right)$$

is such that

$$\left\| W\pi^M U_M^S - i_{S_K}^*(W\pi^M) U^{\widehat{F_K}} \right\| \leq \frac{\epsilon C \left\| W\pi^M \right\|_2}{4(1 - \lambda^2)}$$

What does ML give to MF?

- numerical evaluation of almost any thought experiment becomes feasible: we see 'solutions' of problems numerically which we have never seen before.
- new concepts of modeling.

What does ML give to MF?

- numerical evaluation of almost any thought experiment becomes feasible: we see 'solutions' of problems numerically which we have never seen before.
- new concepts of modeling.

What does MF give to ML?

- it is important to mathematize problems to support understanding.
- roles of randomness.

What does MF give to ML?

- it is important to mathematize problems to support understanding.
- roles of randomness.

Summary and Outlook

- construct universal and randomized reservoirs beyond (branched) rough paths, for instance in the realm of regularity structures.
- improve the JL argument and adapt it to algebraic structures.

Summary and Outlook

- construct universal and randomized reservoirs beyond (branched) rough paths, for instance in the realm of regularity structures.
- improve the JL argument and adapt it to algebraic structures.

References

- H. Bühler, L. Gonon, J. Teichmann, and B. Wood: *Deep Hedging*, Arxiv, 2018.
- C. Cuchiero, M. Larsson, J. Teichmann: *Controlled neural ordinary differential equations*, working paper, 2019.
- C. Cuchiero, L. Gonon, L. Grigoryeva, J.-P. Ortega, J. Teichmann: *Representation of Dynamics by randomized signatures*, working paper, 2019.
- T. Lyons, *Rough paths, Signatures and the modelling of functions on streams Terry Lyons*, Arxiv, 2014.