Randomized signature and reservoir computing *(RSS Applied Probability Section: Rough path theory in machine learning)*

Josef Teichmann

ETH Zürich

August 26, 2020

Introduction

Goal of this talk is ...

- to present the paradigm of reservoir computing and connect it to rNNs and signature representations.
- to apply random projection techniques 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, 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 representations.
- to apply random projection techniques 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, 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 representations.
- to apply random projection techniques 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, Martin Larsson, and Juan-Pablo Ortega)

CODE

We consider differential equations of the form

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

to construction evolutions in state space E (could be a manifold of finite or infinite dimension) depending on local characteristics, initial value $y \in E$ and the control u.

If the map $y \to Y_T$ is considered CODEs are an exciting model for feedforward neural networks, residual networks, etc (see joint work with Christa Cuchiero and Martin Larsson).

CODEs: control as input

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

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

and 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.

It can be 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 dynamic, e.g. a time series. An example: learn a given evolution on state space E:

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 are CODEs/RDEs

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 manifold 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 $x \in E$

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

with remainder term

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

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 $e_1, \dots, e_d$ (actually a Hopf Alebra). A typical element $a \in \mathbb{A}_d$ is written as

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

sums and products are defined in the natural way. We consider the complete locally convex topology making all projections $a \mapsto a_{i_1 \dots i_k}$ 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*

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

Actually $\mathrm{Sig}(e)$ *takes values in a Lie group G and any element of G can be reached up to arbitrary order of accuracy by such evolutions starting at e. Additionally the restriction of linear maps on G is an algebra.*

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 a 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 a 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 a 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 a 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 a 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

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

We use the convention $\inf \emptyset = +\infty$ since the function is only finite on $\operatorname{span}(Q)$. 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 an 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

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

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*

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

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

Hence $f^*(\text{r-Sig})$ *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.

An alternative perspective

Instead of applying the JL Lemma directly on $\mathbb{A}_d$ we could construct faithful representations and evaluate them. Consider a manifold M and $V_1, \ldots, V_d$ vector fields on M such that the map

$$e_i \mapsto V_i$$

from the Lie algebra $\mathfrak{g} \subset \mathbb{A}_d$ to the Lie algebra of vector fields does not have a kernel, in other words there are no non-trivial relations among Lie brackets of the vector fields $V_1, \ldots, V_d$. Then the algebra of (formal) differential operators generated by $V_1, \ldots, V_d$ and $\mathbb{A}_d$ are isomorphic.

An alternative perspective

Furthermore the solution of the transport equation

$$df_t(x) = \sum_{i=1}^{d} V_i f_t(x) du^i(t)$$

and signature have the same expressive power. Notice that $f_t(x) = f(X_t)$ where

$$dX_t = \sum_{i=1}^{d} V_i(X_t) du^i(t), X_0 = x$$

for $x \in M$, $f \in C^\infty(M)$.

An alternative perspective

This yields an alternative perspective to understanding reservoirs constructed by generic vector fields: consider random vector fields, such that they are generic, i.e. without non-trivial relations, consider random smooth functions on M and randomly chosen points $x \in M$, then the vector $(f_t(x))_{0 \leq t \leq T}$ of paths approximates signature up to arbitrary precision. This construction can be fully parallelized and does only depend on a low dimensional evaluation of the above CODE

$$dX_t = \sum_{i=1}^{d} V_i(X_t) du^i(t), X_0 = x$$

for $x \in M$ and $f \in C^\infty(M)$.

An example from Finance: learn the dynamics of SP500

We assume that a traded quantify (we neglect interest rates here) follows an unknown high-dimensional Ito diffusion

$$dY_t = V(Y_t)dt + \sum_{i=1}^{d} V_i(Y_t)dB_t^i \, .$$

No arbitrage theory suggests that there is actually an equivalent measure change on Wiener space such that Y is a local martingale, i.e. there exists a market price of risk (which is of course not observable path wise).

An example from Finance: learn the dynamics of SP500

Still we are able to write

$$dY_t = \sum_{i=1}^{d} V_i(Y_t) dM_t^i \,,$$

where M is a Brownian motion with drift. Under mild assumptions on the vector fields we are able to reconstruct M up to orthogonal transformations from Y in a pathwise manner, i.e. we have M and Y at hand. Then we can learn the still unknown dynamics of Y via RC via regression. With an estimator for the market price of risk, the calibrated model can be used for predictions and pricing.

References

- C. Cuchiero, M. Larsson, J. Teichmann: *Controlled neural ordinary differential equations*, accepted, SIAM, 2020.
- C. Cuchiero, L. Gonon, L. Grigoryeva, J.-P. Ortega, J. Teichmann: *Representation of Dynamics by randomized signatures*, working paper, 2020.
- T. Lyons, *Rough paths, Signatures and the modelling of functions on streams Terry Lyons*, Arxiv, 2014.