A representation theoretic view on signature transforms

Josef Teichmann
(based on several joint works)

ETHZ

Fields Institute September 2022

Goal of the talk

- A bridge between signature transforms on path space and features created by recurrent neural networks (fully trained or in the sense of reservoir computing) might be given by randomized signature.
- We provide a representation theoretic viewpoint on randomized signature.
- We show an example of learning dynamics, where randomized signatures, neural networks (and a bit of domain knownledge) are used to optimally reconstruct dynamics.

Goal of the talk

- A bridge between signature transforms on path space and features created by recurrent neural networks (fully trained or in the sense of reservoir computing) might be given by randomized signature.
- We provide a representation theoretic viewpoint on randomized signature.
- We show an example of learning dynamics, where randomized signatures, neural networks (and a bit of domain knownledge) are used to optimally reconstruct dynamics.

Goal of the talk

- A bridge between signature transforms on path space and features created by recurrent neural networks (fully trained or in the sense of reservoir computing) might be given by randomized signature.
- We provide a representation theoretic viewpoint on randomized signature.
- We show an example of learning dynamics, where randomized signatures, neural networks (and a bit of domain knownledge) are used to optimally reconstruct dynamics.

Signature in a nutshell - notation

- The signature takes values in the free algebra generated by d indeterminates $e_1, \dots, e_d$ given by

$$T((\mathbb{R}^d)) := \{a = \sum_{k=0}^{\infty} \sum_{i_1,\dots,i_k=1}^{d} a_{i_1 \dots i_k} e_{i_1} \cdots e_{i_k}\}.$$

Signature in a nutshell - notation

- The signature takes values in the free algebra generated by d indeterminates $e_1, \dots, e_d$ given by

$$T((\mathbb{R}^d)) := \{a = \sum_{k=0}^{\infty} \sum_{i_1,\dots,i_k=1}^{d} a_{i_1 \dots i_k} e_{i_1} \cdots e_{i_k}\}.$$

- 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.

Signature in a nutshell - definition

Signature of u is the unique solution of the following CODE in $T((\mathbb{R}^d))$

$$d\,\mathrm{Sig}_{s,t} = \sum_{i=1}^{d} \mathrm{Sig}_{s,t}\, e_i du_t^i, \quad \mathrm{Sig}_{s,s} = 1.$$

and is apparently given by

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

Signature and its connection to reservoir computing

The following "splitting theorem" is the precise link to reservoir computing. We suppose here that the controlled ordinary differential equation with characteristics $V_1, \dots, V_d$ admits a unique global solution given by an smooth evolution operator Evol such that $Y_t = \mathrm{Evol}_t(y)$.

Theorem

Let Evol *be a smooth evolution operator on* $\mathbb{R}^N$ *such that* $(\mathrm{Evol}_t(y))_t$ *satisfies a controlled ordinary differential equation with characteristics* $V_1, \dots, V_d$. *Then for any smooth function* $g : \mathbb{R}^N \to \mathbb{R}$ *and for every* $M \geq 0$ *there is a time-homogenous linear map* W *depending on* $(V_1, \dots, V_d, g, M, y)$ *from* $T^M(\mathbb{R}^d) \to \mathbb{R}$ *such that*

$$g\Big(\mathrm{Evol}_t(y)\Big) = W\Big(\pi_M(\mathrm{Sig}_t)\Big) + \mathcal{O}\Big(t^{M+1}\Big),$$

where $\pi_M : T((\mathbb{R}^d)) \to T^M(\mathbb{R}^d)$ *is the canonical projection.*

Remark

For the proof see e.g. Lyons (1998). It can however be proved in much more generality, e.g. on convenient vector spaces.

Is signature a good reservoir?

- This split is not yet fully in spirit of reservoir computing, since unlike a physical systems where the evaluations are ultrafast, computing signature up to a high order can take a while, in particular if d is large.
- Moreover, regression on signature is the analog on path space of a polynomial approximation, which can have several disadvantages.
- Remedy: information compression by Johnson-Lindenstrauss projection.

The Johnson-Lindenstrauss (JL) lemma

We here state the classical version of the Johnson-Lindenstrauss Lemma.

Lemma

For every $0 < \epsilon < 1$, *and every* Q *consisting of* N *point set in some* $\mathbb{R}^n$, *there is a linear map* $f : \mathbb{R}^n \to \mathbb{R}^k$ *with* $k \geq \frac{24 \log N}{3\epsilon^2 - 2\epsilon^3}$ *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 $v_1, v_2 \in Q$, *i.e. the geometry of* Q *is almost preserved after the projection.*

- The map f is called (JL) map and it can be drawn randomly from a set of linear projection maps.
- Indeed, take a $k \times n$ matrix A of with iid standard normal entries. Then $\frac{1}{\sqrt{k}} A$ satisfies the desired requirements with high probability .
- We apply this remarkable result to obtain "versions of signature" in lower dimensional spaces.

Towards randomized signature

We look for (JL) maps on $T^M(\mathbb{R}^d)$ which preserve its geometry encoded in some set of (relevant) directions Q. In order to make this program work, we need the following definition:

Definition

Let Q be any (finite or infinite) set of elements of norm one in $T^M(\mathbb{R}^d)$ with $Q = -Q$. For $v \in T^M(\mathbb{R}^d)$ we define the function

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

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

- The function $\|.\|_Q$ behaves precisely like a norm on the span of Q if $\sup\{\|v\| \mid v \in Q\} < \infty$.
- Additionally $\|v\|_{Q_1} \geq \|v\|_{Q_2}$ for $Q_1 \subset Q_2$.

Towards randomized signature - a first estimate

Proposition

Fix $M \geq 1$ and $\epsilon > 0$. Moreover, let Q be any N point set of vectors with norm one in $T^M(\mathbb{R}^d)$. Then there is linear map $f : T^M(\mathbb{R}^d) \to \mathbb{R}^k$ (with k being the above JL constant with N), such that

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

for all $v_1, v_2 \in \mathrm{span}(Q)$, where $f^* : \mathbb{R}^k \to T^M(\mathbb{R}^d)$ denotes the adjoint map of f with respect to the standard inner product on $\mathbb{R}^k$.

Towards randomized signature - a first estimate

Proposition

Fix $M \geq 1$ and $\epsilon > 0$. Moreover, let Q be any N point set of vectors with norm one in $T^M(\mathbb{R}^d)$. Then there is linear map $f : T^M(\mathbb{R}^d) \to \mathbb{R}^k$ (with k being the above JL constant with N), such that

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

for all $v_1, v_2 \in \mathrm{span}(Q)$, where $f^* : \mathbb{R}^k \to T^M(\mathbb{R}^d)$ denotes the adjoint map of f with respect to the standard inner product on $\mathbb{R}^k$.

- By means of this special JL map associated to a point set Q we can now "project signature" without loosing too much information.
- We can then solve the projected and obtain – up to some time – a solution which is ϵ-close to signature.
- By a slight abuse of notation we write Sig_t for the truncated version $\pi_M(\mathrm{Sig}_t)$ in $T^M(\mathbb{R}^d)$.

Randomized signature is as expressive as signature

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

Let u be a smooth control and f a JL map from $T^M(\mathbb{R}^d) \to \mathbb{R}^k$ where k is determined via some fixed ϵ and a fixed set Q. We denote by r-Sig *the smooth evolution of the following controlled differential equation on $\mathbb{R}^k$*

$$dX_t = \sum_{i=1}^{d} \Big(\frac{1}{\sqrt{n}} f(f^*(X_t)e_i) + (1 - \frac{1}{\sqrt{n}}) f(\mathrm{Sig}_t\, e_i) \Big) du^i(t)\,,\ X_0 \in \mathbb{R}^k\,,$$

where $n = \dim(T^M(\mathbb{R}^d))$. Then for each $w \in T^M(\mathbb{R}^d)$

$$|\langle w, \mathrm{Sig}_t - f^*(\text{r-Sig}_t(X_0))\rangle|$$

$$\leq \big|\langle w, \mathrm{Evol}_t(1 - f^*(X_0))\rangle\big| + C\epsilon \sum_{i=1}^{d} \int_0^t \|\,\mathrm{Evol}_r^*\, w\|_Q \|\,\mathrm{Sig}_r\, e_i\|_Q\, dr\,,$$

where Evol *denotes here the evolution operator corresponding to $dZ_t = \sum_{i=1}^d \frac{1}{\sqrt{n}}(f^* \circ f)(Z_t e_i) du^i(t)$ and $C = \sup_{s \leq r \leq t,\, i} \left| \frac{du^i(r)}{dr} \right|$.*

r-Sig as random dynamical system

We can actually calculate approximately the vector fields which determine the dynamics of r-Sig by generic random elements.

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

For $M \to \infty$ (and thus $n \to \infty$) the entries of the linear maps

$$y \mapsto \frac{1}{\sqrt{n}} f(f^*(y)e_i)$$

for $i = 1, \dots, d$, are asymptotically normally distributed with independent entries. The time dependent bias terms

$$(1 - \frac{1}{\sqrt{n}}) f(\mathrm{Sig}_t\, e_i)$$

are as well asymptotically normally distributed with independent entries.

Randomized signature as reservoir

Practical implementation of randomized signature

- Given a set of hyper-parameters $\theta \in \Theta$, and a dimension k, choose randomly (often just by independently sampling from a normal distribution) matrices $A_1, \dots, A_d \in \mathbb{R}^{k \times k}$ as well as (bias) vectors $b_1, \dots, b_d$.
- Then one can tune the hyper-parameters and the dimension k such that

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

approximates the CODE Y locally in time via a linear readout W up to arbitrary precision.

- The process X will serve as reservoir. Note that again it does not depend on the specific dynamics of Y which should be learned.

Representation theory

Instead of applying the JL Lemma directly on $\mathbb{A}_d$ we could construct faithful representations and evaluate them. Consider a manifold M and $V_1, \dots, V_d$ 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 $V_1, \dots, V_d$. Then the algebra of (formal) differential operators generated by $V_1, \dots, V_d$ and $\mathbb{A}_d$ are isomorphic.

Controlled transport PDEs and the method of characteristics

Furthermore the solution of the transport equation

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

and signature have the same expressive power.

Notice that $f_t(s,x) = f(X_t)$ where

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

for $x \in M$, $f \in C^\infty(M)$ and $s \in \mathbb{R}$.

Generic controlled transport PDEs represent signatures

This yields an alternative perspective to understanding reservoirs constructed by generic vector fields: consider generic vector fields, then the vector $(f_t(x))_{0\leq t\leq T}^{f,x}$ of paths for approximate signature up to arbitrary precision.

This construction can be fully parallelized and does only depend on a low dimensional evaluation of the above CODE

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

for $x \in M$, $f \in C^\infty(M)$ and $s \in \mathbb{R}$.

A strengthened version

Let σ be a random (i.e. all coefficients are independently randomly chosen), real analytic activation function and $A_1, \dots, A_d$, $b_1, \dots, b_d$ be randomly chosen matrices (each with absolutely continuous laws with respect to Lebesgue measure) and vectors of dimension k, then

$$V_i(x) := \sigma(A_i x + b_i)$$

defines vector fields on $\mathbb{R}^k$, which satisfy an even stronger condition, namely that

$$\sum_{i_1,\dots,v_k=1, k\leq m}^{d} b_{i_1\dots i_k} V_{i_1} \cdots V_{i_k} \,\mathrm{id}(x) = 0$$

for all $x \in \mathbb{R}^k$ implies that all coefficients vanish.

Then the flow at time t

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

for all $x \in \mathbb{R}^k$ and s contains the same information as signature at t.

Proof

Generic vector fields are of course meager in the set set of all vector fields, but can also be constructed by random procedures. This can be seen by using appropriate metrics on real analytic functions. This does not allow to conclude about random constructions but gives a hint that real analytic functions satisfying polynomial differential relations are 'small'.

Proof

Let us first construct a generic, randomly chosen activation function σ:

There is an activation function σ, i.e. a real analytic, non-constant and bounded function such that there is no polynomial in finitely many variables from $\sigma^{(n)}(z_{il})$, for $d \times k$ commuting indeterminates z_{il}, which vanishes (in other words σ does not satisfy any polynomial differential relations). In other words: $\sigma^{(n)}(z_{il})$ behaves like a triple-indexed, countable system of free, commutative variables.

For the proof it is sufficient to understand that any polynomial relation enforces σ to lie in a subset of analytic functions which has measure zero with respect to independent variations of all coefficients of σ. One can order the polynomial relations by the maximal degree of derivative in σ and the overall polynomial degree and therefore proves the set of all functions satisfying *any* polynomial relation being negligible.

Proof

Let σ be a generic activation function and consider the vector of formal powerseries in commuting indeterminates z_{il}

$$V_i \,\mathrm{id}(x) := (\sigma(z_{il}))_{1 \le l \le k}$$

for $i = 1, \ldots, d$. We define (as formal power series)

$$V_i V_j \,\mathrm{id}(x) := (\sum_{l=1}^{k} a_{jl} \sigma(z_{il}) \sigma^{(1)}(z_{jm}))_{1 \le m \le k}$$

for invertible $k \times k$ matrices (a_{jl}) yielding a non-commutative algebra generated by V_i, $1 \le i \le d$.

It is then clear by induction that

$$\sum_{i_1, \ldots, v_k = 1, k \le m}^{d} b_{i_1 \cdots i_k} V_{i_1} \cdots V_{i_k} \,\mathrm{id}(x) = 0$$

leads to all coefficients vanishing.

Proof

The very reason is that we can identify by looking at the highest derivative of σ the vector fields applied first to id. This yields by the previously established freeness of $\sigma^{(n)}(z_{il})$ and induction the result.

Randomized Signature

For fixed random matrices $A,1,\dots,A_d$ (components indedently sampled with respect to a law absolutely continuous to Lebesgue measure) and fixed x the dynamical systems

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

for all possible choices of b_i is as expressive as signature.

An application: path dependent neural jump ODEs

The observer does *not* have knowledge on the specific dynamics of X, but she knows sufficiently many paths of X and of the corresponding observations to allow for approximations of the conditional expectation.

Additionally it is assumed that between two observation points, i.e. when the information σ algebra does not grow, the conditional expectation is absolutely continuous with respect to time.

Goal is to parametrize the respective path space functionals and to provide loss functions for learning.

Related Work

(Recurrent) neural networks, signature transforms and the neural ODE framework are the main ingredients for the (path-dependent) NJ-ODE model.

The first works, in which they were combined to a model similar to the one we use, are by Yulia Rubanova et al (Neurips 2019) and Edward De Brouwer et al (Neurips 2019) (the so called Gated Recurrent Unit ODE model). We have a different setup in view of model and objective function, we provide convergence guarantees, and in particular a fully general stochastic setup for the observed process. Main role model for this paper is Calypso Herrera et al (ICLR 2021), where the Markovian case is treated.

The most related work in the context of the labelling problem, besides GRU-ODEs, is the neural controlled differential equation (NCDE) by Patrick Kidger et al.

Detailed assumptions

Let $d_X \in \mathbb{N}$ and $T > 0$ be the fixed time horizon. Consider a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F} := \{\mathcal{F}_t\}_{0 \le t \le T}, \mathbb{P})$, on which an adapted càdlàg stochastic process[1] $X := (X_t)_{t \in [0,T]}$ taking values in $\mathbb{R}^{d_X}$. We define the running maximum process

$$X_t^\star := \sup_{0 \le s \le t} |X_s|, \quad 0 \le t \le T.$$

Moreover, let $\mathcal{J}$ be the random set of discontinuity times of X, defined for every $\omega \in \Omega$ as $\mathcal{J}(\omega) := \{t \in [0, T] | \Delta X_t(\omega) \neq 0\}$.

[1] A stochastic process is a collection of random variables $X_t : \Omega \to \mathbb{R}^{d_X}, \omega \mapsto X_t(\omega)$ for $0 \le t \le T$.

Detailed assumptions

We considering another probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$, on which the random observation times of the stochastic process are defined. In particular, we define

- $n : \tilde{\Omega} \to \mathbb{N}_{\geq 0}$, a random variable with $\mathbb{E}_{\tilde{\mathbb{P}}}[n] < \infty$, which is the random number of observations, and
- $t_i : \tilde{\Omega} \to [0, T]$ for $0 \leq i \leq n$, the *sorted* [2] random variables, describing the random observation times.

[2] For all $\tilde{\omega} \in \tilde{\Omega}$, $0 = t_0 < t_1(\tilde{\omega}) < \cdots < t_{n(\tilde{\omega})}(\tilde{\omega}) \leq T$.

Detailed assumptions

Moreover, we let $K := \max\left\{k \in \mathbb{N} \,|\, \tilde{\mathbb{P}}(n \geq k) > 0\right\} \in \mathbb{N} \cup \{\infty\}$ be the maximal value of n. We use the notation $\mathcal{B}([0,T])$ for the Borel σ-algebra of the set $[0,T]$ and define for each $1 \leq k \leq K$

$$\lambda_k : \mathcal{B}([0,T]) \to [0,1], \quad B \mapsto \lambda_k(B) := \frac{\tilde{\mathbb{P}}(n \geq k, (t_k -) \in B)}{\tilde{\mathbb{P}}(n \geq k)},$$

which is a probability measure on the time interval $[0,T]$. The time of the last observation before a certain time t is defined as

$$(t, \tilde{\omega}) \mapsto \tau(t, \tilde{\omega}) := \max\{t_i(\tilde{\omega}) | 0 \leq i \leq n(\tilde{\omega}), t_i(\tilde{\omega}) \leq t\}.$$

Detailed assumptions

The observation mask $M = (M_k)_{0 \leq k \leq K}$ is a sequence of random variables on $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ taking values in $\{0,1\}^{d_X}$ such that the j-th coordinate of the k-th element of the sequence, i.e. $M_{k,j}$, signals whether the j-th coordinate of the stochastic process, denoted $X_{t_k,j}$, is observed at observation time t_k. In particular, $M_{k,j} = 1$ means that it is observed, while $M_{k,j} = 0$ means that it is not. By abuse of notation, we also write $M_{t_k} := M_k$.

Information σ-algebra

We define the filtration of the currently available information $\mathbb{A} := (\mathcal{A}_t)_{t\in[0,T]}$ by

$$\mathcal{A}_t := \boldsymbol{\sigma}\left(X_{t_i,j} | t_i \leq t,\, j \in \{1 \leq l \leq n | M_{t_i,l} = 1\}\right),$$

where t_i are the observation times and $\boldsymbol{\sigma}(\cdot)$ denotes the generated σ-algebra. By the definition of τ we have $\mathcal{A}_t = \mathcal{A}_{\tau(t)}$ for all $t \in [0, T]$.

$(\Omega \times \tilde{\Omega}, \mathcal{F} \otimes \tilde{\mathcal{F}}, \mathbb{F} \otimes \tilde{\mathcal{F}}, \mathbb{P} \times \tilde{\mathbb{P}})$ is the filtered product probability space which, intuitively speaking, combines the randomness of the stochastic process with the randomness of the observations.

Assumptions on X

We denote the conditional expectation process of X by $\hat{X} = (\hat{X}_t)_{0 \leq t \leq T}$, defined by $\hat{X}_t := \mathbb{E}_{\mathbb{P} \times \tilde{\mathbb{P}}}[X_t | \mathcal{A}_t]$ and remark that $\hat{X}_{\tau(t)} \neq X_{\tau(t)}$ in general, since observations might be incomplete. Moreover, we define for any $0 \leq t \leq T$ the process $\tilde{X}^{\leq t}$ to be a continuous version of the rectilinear interpolation of the observations of X until time t.

Assumptions on X

Its j-th coordinate at time $0 \le s \le T$ is given by

$$\tilde{X}^{\le t}_{s,j} := \begin{cases} X_{t_{l(s,t)},j} \frac{t_{\ell(s,t)} - s}{t_{\ell(s,t)} - t_{\ell(s,t)-1}} + X_{t_{\ell(s,t)},j} \frac{s - t_{\ell(s,t)-1}}{t_{\ell(s,t)} - t_{\ell(s,t)-1}}, & \text{if } t_{\ell(s,t)-1} < s \le t_{\ell(s,t)}, \\ X_{t_{l(s,t)},j}, & \text{if } s \le t_{\ell(s,t)-1}, \end{cases}$$

where

$$l(s,t) := l(s,t,j) := \max\{0 \le l \le n | t_l \le \min(s,t), M_{t_l,j} = 1\},$$
$$\ell(s,t) := \ell(s,t,j) := \inf\{1 \le \ell \le n | s \le t_\ell \le t, M_{t_\ell,j} = 1\},$$

with the standard definition that the infimum of the empty set is ∞ and the additional definition that $t_\infty := T$.

Assumptions on X

In particular, $\tilde{X}^{\leq t}$ is the rectilinear interpolation (sometimes denoted as forward-fill), except that its jumps at $t_{l(s,t)}$ are replaced by linear interpolations between the previous observation time $t_{\ell(s,t)-1}$ and $t_{l(s,t)}$. It is important to note, that this is not solely a coordinate-wise interpolation, since the given coordinate might not have been observed at the previous observation time. Moreover, by this definition, $\tilde{X}^{\leq \tau(t)}$ is $\mathcal{A}_{\tau(t)}$-measurable for all t, and for any $r \geq t$ and all $s \leq \tau(t)$ we have $\tilde{X}_s^{\leq \tau(t)} = \tilde{X}_s^{\leq \tau(r)}$.

Assumptions on X

Using $\mathcal{A}_t = \mathcal{A}_{\tau(t)}$ and that $\tilde{X}^{\leq \tau(t)} \in \mathcal{A}_{\tau(t)}$ carries all information available in $\mathcal{A}_{\tau(t)}$, we know that there exist measurable functions $F_j : [0,T] \times [0,T] \times BV^c([0,T]) \to \mathbb{R}$ such that $\hat{X}_{t,j} = F_j(t, \tau(t), \tilde{X}^{\leq \tau(t)})$.

Assumptions on X

Assumption

We assume that

1. *for every* $1 \leq k, l \leq K$, M_k *is independent of* t_l *and of* n, $\tilde{\mathbb{P}}(M_{k,j} = 1) > 0$ *and* $M_{0,j} = 1$ *for all* $1 \leq j \leq d_X$ *(i.e. every coordinate can be observed at any observation time and* X *is completely observed at* 0*) and* $|M_k|_1 > 0$ *for every* $1 \leq k \leq K$ $\tilde{\mathbb{P}}$*-almost surely (i.e. at every observation time at least one coordinate is observed),*
2. *the probability that any two observation times are closer than* $\epsilon > 0$ *converges to* 0 *when* ϵ *does, i.e. if* $\delta(\tilde{\omega}) := \min_{0 \leq i \leq n(\tilde{\omega})} |t_{i+1}(\tilde{\omega}) - t_i(\tilde{\omega})|$ *then* $\lim_{\epsilon \to 0} \tilde{\mathbb{P}}(\delta < \epsilon) = 0$,
3. *almost surely* X *is not observed at a jump, i.e.* $(\mathbb{P} \times \tilde{\mathbb{P}})(t_j \in \mathcal{J} | j \leq n) = (\mathbb{P} \times \tilde{\mathbb{P}})(\Delta X_{t_j} \neq 0 | j \leq n) = 0$ *for all* $1 \leq j \leq K$,
4. F_j *are continuous and differentiable in their first coordinate* t *such that their partial derivatives with respect to* t*, denoted by* f_j*, are again continuous and there exists a* $B > 0$ *and* $p \in \mathbb{N}$ *such that for every* $t \in [0, T]$ *the functions* f_j, F_j *are polynomially bounded in* $X^\star$*, i.e.*

$$|F_j(\tau(t), \tau(t), \tilde{X}^{\leq \tau(t)})| + |f_j(t, \tau(t), \tilde{X}^{\leq \tau(t)})| \leq B(X_t^\star + 1)^p,$$

5. $X^\star$ *is* L^{2p}*-integrable, i.e.* $\mathbb{E}[(X_T^\star)^{2p}] < \infty$.

Detailed assumptions on X

Under Assumption 1 we can rewrite $\hat{X}$ by the fundamental theorem of calculus as

$$\hat{X}_{t,j} = F_j(\tau(t), \tau(t), \tilde{X}^{\leq \tau(t)}) + \int_{\tau(t)}^{t} f_j(s, \tau(t), \tilde{X}^{\leq \tau(t)}) ds,$$

implying that it is càdlàg. We remark that jumps of $\hat{X}$ occur only at new observation times, i.e., at t_i, for $1 \leq i \leq n$.

The model

Note that we can not use the signature of the true path $(X_s)_{0\leq s\leq t}$ of the data up to time t as input, since we only have discrete observations of X at the observation times t_i (which is not sufficient to calculate the signature of X). Instead, we use the shifted interpolation $\tilde{X}^{\leq t} - X_0 \in BV_0^c([0,T])$ up to time t and compute the truncated signature $\pi_n(\tilde{X}^{\leq t} - X_0)$. This signature together with the starting point X_0 include all available information (while the signature of $(X_s)_{0\leq s\leq t}$ would include much more then the available information, i.e., it is not $\mathcal{A}_t$-measurable). Moreover, the interpolation $\tilde{X}^{\leq t}$ has bounded variation, no matter whether this is true for the original path X or not.

The model

Definition

The *path-dependent neural jump ODE (PD-NJ-ODE)* model (of order $n \in \mathbb{N}$) is given by

$$\begin{aligned} H_0 &= \rho_{\theta_2}\left(0, 0, \pi_n(0), X_0\right), \\ dH_t &= f_{\theta_1}\left(H_{t-}, t, \tau(t), \pi_n(\tilde{X}^{\leq \tau(t)} - X_0), X_0\right) dt \\ &\quad + \left(\rho_{\theta_2}\left(H_{t-}, t, \pi_n(\tilde{X}^{\leq \tau(t)} - X_0), X_0\right) - H_{t-}\right) du_t, \\ Y_t &= g_{\theta_3}(H_t). \end{aligned} \tag{1}$$

The functions $f_{\theta_1}, \rho_{\theta_2}$ and g_{θ_3} are feedforward neural networks with parameters $\theta = (\theta_1, \theta_2, \theta_3) \in \Theta$ and u is the jump process counting the observations.

The objective function

Let $\mathbb{D}$ to be the set of all $\mathbb{R}^{d_X}$-valued $\mathbb{A}$-adapted processes on the probability space $(\Omega \times \tilde{\Omega}, \mathcal{F} \otimes \tilde{\mathcal{F}}, \mathbb{F} \otimes \tilde{\mathcal{F}}, \mathbb{P} \times \tilde{\mathbb{P}})$. Then we define our objective functions

$$\Psi : \mathbb{D} \to \mathbb{R},$$

$$Z \mapsto \Psi(Z) := \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^{n}\left(|M_i \odot (X_{t_i} - Z_{t_i})|_2 + |M_i \odot (Z_{t_i} - Z_{t_{i-}})|_2\right)^2\right], \tag{2}$$

$$\Phi : \Theta \to \mathbb{R}, \theta \mapsto \Phi(\theta) := \Psi(Y^\theta(X)), \tag{3}$$

where $\odot$ is the element-wise multiplication (Hadamard product) and Φ will be our (theoretical) loss function. Remark that from the definition of Y^θ it directly follows that it is an element of $\mathbb{D}$, hence Φ is well-defined.

The objective function

Let us assume, that we observe $N \in \mathbb{N}$ independent realisations of the path X together with independent realizations of the observation mask M at times $(t_1^{(j)}, \cdots, t_{n^j}^{(j)})$, $1 \leq j \leq N$, which are themselves independent realisations of the random vector $(n, t_1, \cdots, t_n)$. In particular, let us assume that $X^{(j)} \sim X$, $M^{(j)} \sim M$ and $(n^j, t_1^{(j)}, \cdots, t_{n^j}^{(j)}) \sim (n, t_1, \cdots, t_{n^j})$ are i.i.d. random processes (respectively variables) for $1 \leq j \leq N$ and that our training data is one realisation of them.

We write $Y^{\theta,j} := Y^\theta(X^{(j)})$. Then the Monte Carlo approximation of our loss function

$$\hat{\Phi}_N(\theta) := \frac{1}{N} \sum_{j=1}^{N} \frac{1}{n^j} \sum_{i=1}^{n^j} \left(\left| M_i^{(j)} \odot \left(X_{t_i^{(j)}}^{(j)} - Y_{t_i^{(j)}}^{\theta,j} \right) \right|_2 + \left| M_i^{(j)} \odot \left(Y_{t_i^{(j)}}^{\theta,j} - Y_{t_i^{(j)}-}^{\theta,j} \right) \right|_2 \right)^2 \tag{4}$$

converges $(\mathbb{P} \times \tilde{\mathbb{P}})$-a.s. to $\Phi(\theta)$ as $N \to \infty$, by the law of large numbers.

Monte Carlo approximation results

Theorem

Let $\theta_m^{\min} \in \Theta_m^{\min} := \operatorname{argmin}_{\theta \in \Theta_m}\{\Phi(\theta)\}$ *for every* $m \in \mathbb{N}$. *If Assumption 1 is satisfied, then, for* $m \to \infty$, *the value of the loss function* Φ (3) *converges to the minimal value of* Ψ (2) *which is uniquely achieved by* $\hat{X}$, *i.e.*

$$\Phi(\theta_m^{\min}) \xrightarrow{m\to\infty} \min_{Z\in\mathbb{D}} \Psi(Z) = \Psi(\hat{X}).$$

Furthermore, for every $1 \leq k \leq K$ *we have that* $Y^{\theta_m^{\min}}$ *converges to* $\hat{X}$ *as random variable in* $L^1(\Omega \times [0,T], \mathbb{P} \times \lambda_k)$. *In particular, the limit process* $Y := \lim_{m\to\infty} Y^{\theta_m^{\min}}$ *equals* $\hat{X}$ $(\mathbb{P} \times \lambda_k)$*-almost surely as a random variable on* $\Omega \times [0,T]$.

Monte Carlo approximation results

We now assume the size m of the neural network and of the signature truncation level is fixed and we study the convergence of the Monte Carlo approximation when the number of samples N increases. Moreover, we show that both types of convergence can be combined. We define $\tilde{\Theta}_M := \{\theta \in \Theta_M \mid |\theta|_2 \leq M\}$, which is a compact subspace of Θ_M and recall, that Θ_M in Theorem 7 can be replaced by $\tilde{\Theta}_M$.

Monte Carlo approximation results

Theorem

Let $\theta^{\min}_{m,N} \in \Theta^{\min}_{m,N} := \arg\inf_{\theta\in\tilde{\Theta}_m}\{\hat{\Phi}_N(\theta)\}$ *for every* $m, N \in \mathbb{N}$. *Then, for every* $m \in \mathbb{N}$, $(\mathbb{P}\times\tilde{\mathbb{P}})$*-a.s.*

$$\hat{\Phi}_N \xrightarrow{N\to\infty} \Phi \quad \text{uniformly on } \tilde{\Theta}_m.$$

Moreover, for every $m \in \mathbb{N}$, $(\mathbb{P}\times\tilde{\mathbb{P}})$*-a.s.*

$$\Phi(\theta^{\min}_{m,N}) \xrightarrow{N\to\infty} \Phi(\theta^{\min}_m) \quad \text{and} \quad \hat{\Phi}_N(\theta^{\min}_{m,N}) \xrightarrow{N\to\infty} \Phi(\theta^{\min}_m).$$

In particular, one can define an increasing sequence $(N_m)_{m\in\mathbb{N}}$ *in* $\mathbb{N}$ *such that for every* $1 \le k \le K$ *we have that* $Y^{\theta^{\min}_{m,N_m}}$ *converges to* $\hat{X}$ *for* $m \to \infty$ *as random variable in* $L^1(\Omega\times[0,T], \mathbb{P}\times\lambda_k)$. *In particular, the limit process* $Y := \lim_{m\to\infty} Y^{\theta^{\min}_{m,N_m}}$ *equals* $\hat{X}$ $(\mathbb{P}\times\lambda_k)$*-almost surely as a random variable on* $\Omega\times[0,T]$.

Experiments

Detailed results can be found in our paper on Arxiv `https://arxiv.org/abs/2206.14284` and the code is available on Github.

Observation grids are independently sampled, usually $2 * 10^4$ sample trajectories are used for training.

Uncertainty Estimation: Conditional Variance

We estimate uncertainty in the sense of conditional variance of the observed process:

$$\mathrm{Var}[X_t \mid \mathcal{A}_{\tau(t)}] = \mathbb{E}[X_t^2 \mid \mathcal{A}_{\tau(t)}] - \mathbb{E}[X_t \mid \mathcal{A}_{\tau(t)}]^2. \tag{5}$$

We show results of an experiment for Brownian motion and its square.

Estimated conditional Variance for BM

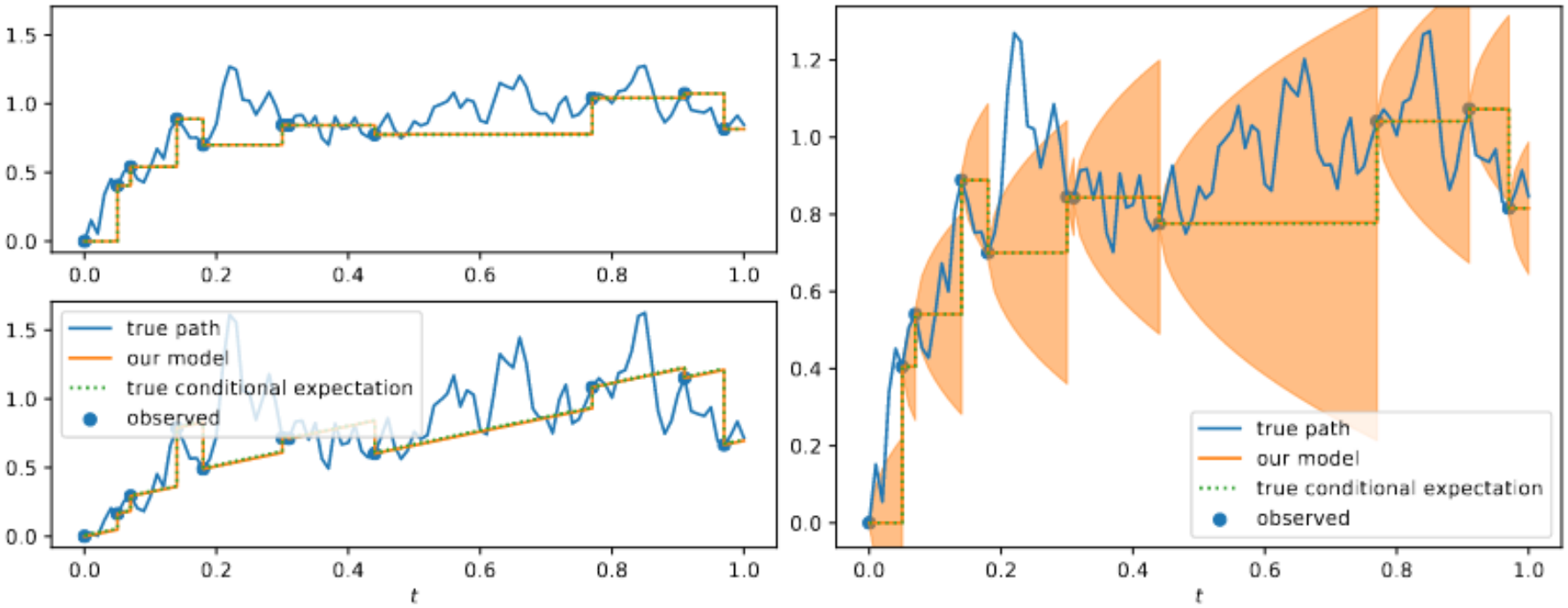


Figure 2: Left: a test sample of a Brownian motion X (top) and its square X^2 (bottom) together with the predicted and true conditional expectation. Right: the same test sample of the Brownian motion X with a confidence interval given as $\hat{\mu}_t \pm \hat{\sigma}_t$.

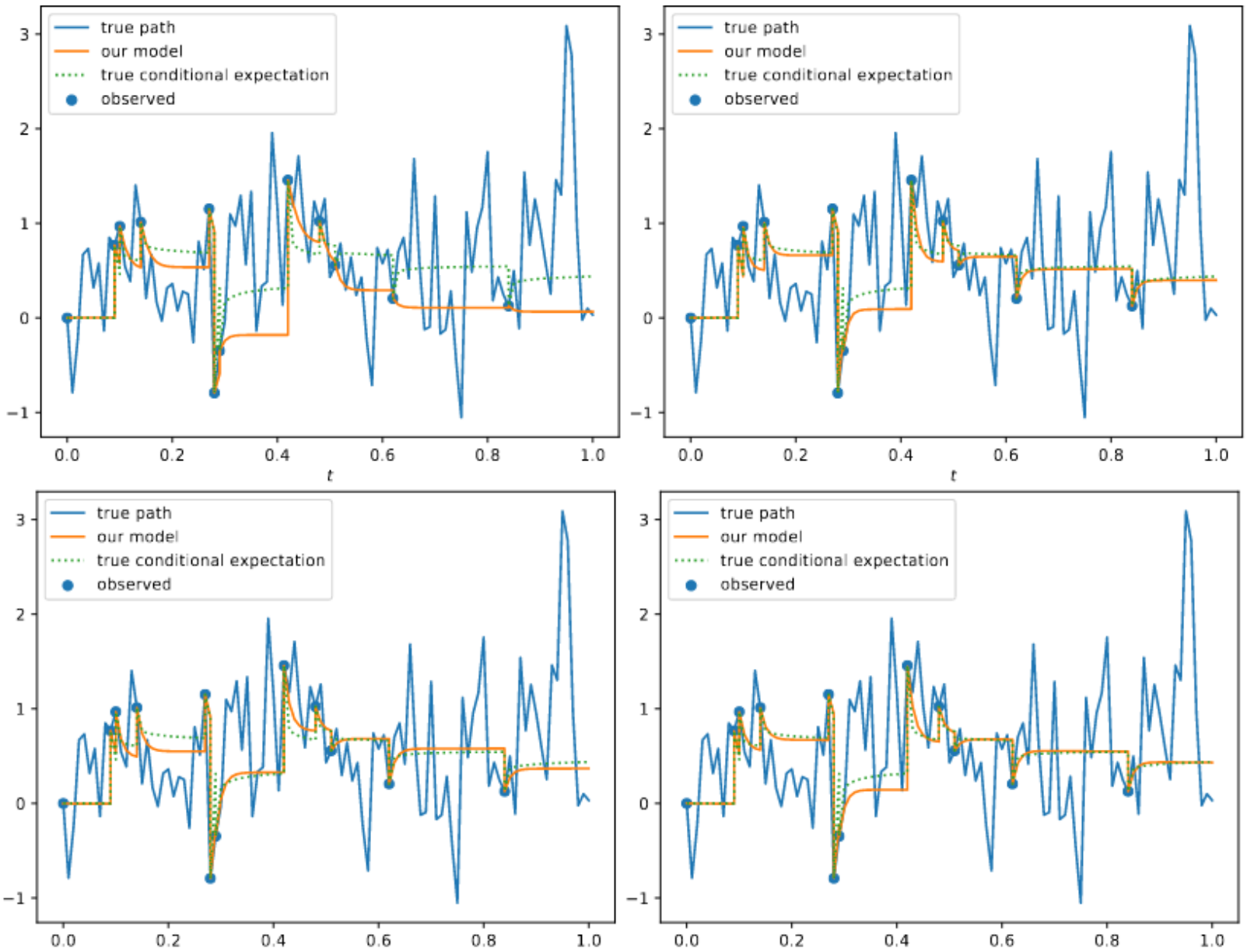
Estimation of fractional BM with $H = 0.05$

In the next example we estimate a fractional Brownian motion irregularly observed. It is remarkable to see that neither truncated signature alone nor recurrence alone are fully able to capture as many features os PD-NJ ODE.

Table 1: Minimal evaluation metrics on the test set of FBM (with $H = 0.05$) within the 200 epochs of training for different NJ-ODE models.

	NJ-ODE	NJ-ODE (with sig.)	NJ-ODE (with RNN)	PD-NJ-ODE
min. evaluation metric	$8.1 \cdot 10^{-2}$	$1.0 \cdot 10^{-2}$	$1.1 \cdot 10^{-2}$	$0.5 \cdot 10^{-2}$

Estimation of fractional BM with $H = 0.05$



Physionet estimation

Table 2: Mean and standard deviation of MSE on the test set of physionet. Results of baselines were reported by Rubanova et al. (2019) and Herrera et al. (2021). Where known, the number of trainable parameters is reported.

	Physionet – MSE ($\times 10^{-3}$)	# params
RNN-VAE	3.055 ± 0.145	-
Latent ODE (RNN enc.)	3.162 ± 0.052	-
Latent ODE (ODE enc)	2.231 ± 0.029	$163'972$
Latent ODE + Poisson	2.208 ± 0.050	$181'723$
NJ-ODE	1.945 ± 0.007	$24'423$
PD-NJ-ODE	$\mathbf{1.930 \pm 0.006}$	$201'691$

LOB ten steps ahead prediction

Table 3: Minimal MSEs (smaller is better) during the training of each model (if applicable) are reported for different LOB datasets.

	BTC	BTC1sec	ETH1sec
last observation	0.11808	1350.44	2.58909
best linear regression	0.12198	1355.04	2.58253
PD-NJ-ODE	**0.11743**	**1343.91**	**2.56636**

Table 4: The true mean value and PD-NJ-ODE's mean predicted value of the midprices 10 steps ahead for the 3 different datasets and their subsets by labels.

	BTC		BTC1sec		ETH1sec	
	true	prediction	true	prediction	true	prediction
overall	0.01143	0.01123	−2.40790	−2.39125	−0.12932	−0.10787
decrease	−0.41470	−0.31175	−47.28358	−28.80446	−2.19563	−1.25667
stationary	0.08016	0.08585	−2.75307	−3.27602	−0.34508	−0.36719
increase	0.44849	0.31417	40.63428	23.33839	2.24031	1.30459

Conclusion

- We provide two view on randomized signature, which constitutes a bridge between RNNs and signature transforms: a compression view and a representation theoretic view.
- This is motivated by paradigms of reservoir computing and widely applied signature methods from rough paths theory.
- We provide a general learning framework for online estimation and prediction of stochastic processes. This also paves a road towards provable machine learning, since we the methodology allows for randomization approaches.