

A Study of Data-driven Market Simulator

Master Thesis

Songyan Hou

Supervisor:

Prof. Dr. Josef Teichmann
Department of Mathematics
ETH Zürich

Submitted: February 27, 2021

Abstract

asd

Contents

| | | |
|----------|---|-----------|
| 1 | Introduction | 3 |
| 2 | Financial times-series stimulation | 4 |
| 2.1 | Challenges in financial scenario | 4 |
| 2.2 | Stylised facts and evaluation of similarity | 4 |
| 2.3 | Classical and modern model | 4 |
| 3 | Signature feature | 5 |
| 3.1 | Signature | 5 |
| 3.1.1 | Signature map | 7 |
| 3.1.2 | Signature stream | 8 |
| 3.2 | Log-signature | 10 |
| 3.2.1 | Lie Series | 11 |
| 3.3 | Universality and characteristics | 12 |
| 3.4 | Discrete signature feature | 15 |
| 3.4.1 | Kernel trick | 16 |
| 3.5 | Low-rank-signature | 17 |
| 4 | Variational Autoencoders | 19 |
| 4.1 | Variational autoencoders | 19 |
| 4.1.1 | Variational inference | 20 |
| 4.1.2 | Reparameterization Trick | 21 |
| 4.2 | Variations of VAE | 22 |
| 5 | Main method | 23 |
| 5.1 | Time series processing | 23 |

| | | |
|----------|---|-----------|
| 5.2 | Generative model with heavy tail | 23 |
| 5.3 | Inverse the log-signature | 23 |
| 5.4 | Evaluation with lifting and normalizing | 23 |
| 6 | Signature stimulation | 24 |
| 6.1 | Controlled stochastic differential equation | 24 |
| 6.2 | Numerical Implementation | 24 |
| 7 | Conclusions and Future Work | 25 |
| 8 | Appendix | 26 |

1. Introduction

2. Financial times-series stimulation

2.1 Challenges in financial scenario

2.2 Stylised facts and evaluation of similarity

2.3 Classical and modern model

3. Signature feature

In this chapter, we give a brief introduction to the signature feature of a path and some variation of signature feature such as log-signature, low-rank-signature and random-signature. Moreover, we discuss the expectation of these signature based features of a stochastic process and emphasize their characteristic ability of the law of stochastic process. For simplicity and readability, we restrict our introduction to the space of continuous bounded variation paths from a compact time interval J to a Euclidean space $E = \mathbb{R}^d$ denoted by $\mathcal{C}_0^1(J, E)$. We postpone the discussion of semi-martingales and rough paths taking values in Banach space E to the appendix and kindly invite interested readers to the comprehensive and rigorous introduction of signature feature in [6] [15].

3.1 Signature

We start the introduction from the tensor algebra in which our signature feature takes value. We consider the non-commuting formal power series of tensors with formal indeterminates as basis of E .

Definition 3.1 (Formal power series). We denote $T((E))$ the space of formal power series of tensors in E that

$$T((E)) = \left\{ \mathbf{a} = (\mathbf{a}_k)_{k \geq 0} : \mathbf{a}_k \in E^{\otimes k} \right\} = \bigoplus_{m \geq 0} E^{\otimes m}$$

endowed with addition and multiplication defined as follows: Let $\mathbf{a} = (\mathbf{a}_k)_{k \geq 0}$, $\mathbf{b} =$

$(\mathbf{b}_k)_{k \geq 0} \in T((E))$ and $\lambda \in \mathbb{R}$. Then

$$\begin{aligned}\mathbf{a} + \mathbf{b} &= (\mathbf{a}_k + \mathbf{b}_k)_{k \geq 0} \\ \mathbf{a} \otimes \mathbf{b} &= \left(\sum_{i+j=k} \mathbf{a}_i \otimes \mathbf{b}_j \right)_{k \geq 0} \\ \lambda \mathbf{a} &= (\lambda \mathbf{a}_k)_{k \geq 0}\end{aligned}$$

Let $T^{(n)}(E)$ denote the truncated tensor algebra space up to order n

$$T^{(n)}(E) = \bigoplus_{k \leq n} E^{\otimes k}$$

Let e_1, \dots, e_d be a finite basis of $E = \mathbb{R}^d$. Then $\mathbf{a} \in T((E))$ has the following linear form

$$\mathbf{a} = \sum_{k \geq 0} \left(\sum_{i_1, \dots, i_k=1}^d a_{i_1, \dots, i_k} e_{i_1} \otimes \dots \otimes e_{i_k} \right), \quad a_{i_1, \dots, i_k} \in \mathbb{R}.$$

and $T^{(n)}(E)$ can be considered as a subspace of $T((E))$.

Definition 3.2. We denote $\mathbf{T}(E)$ the Banach space

$$\mathbf{T}(E) := \left\{ \mathbf{t} \in T((E)) : \|\mathbf{t}\|_{\mathbf{T}(E)} := \sqrt{\sum_{k \geq 0} \|\mathbf{t}_k\|_{E^{\otimes k}}^2} < \infty \right\}.$$

Similarly, we denote $\mathbf{T}^{(n)}(E)$ the truncation of $\mathbf{T}(E)$ up to order n .

Definition 3.3 (Signature). We denote $\mathbf{Sig}_J : \mathcal{C}_0^1(J, E) \rightarrow \mathbf{T}(E)$ the signature map such that for all $X \in \mathcal{C}_0^1(J, E)$

$$\mathbf{Sig}_J(X) = (1, \mathbf{s}_1, \dots) \in \mathbf{T}(E)$$

where

$$\begin{aligned}\mathbf{s}_k &= \int_{t_1 < \dots < t_k \in J} dX_{t_1} \otimes \dots \otimes dX_{t_k} \\ &= \sum_{i_1, \dots, i_k=1}^d \int_{t_1 < \dots < t_k \in J} dX_{t_1}^{i_1} \dots dX_{t_k}^{i_k} \cdot e_{i_1} \otimes \dots \otimes e_{i_k}\end{aligned}$$

Let $\mathbf{Sig}_J^{(n)}$ denote the truncated signature map up to order n

$$\mathbf{Sig}_J^{(n)}(X) = (1, \mathbf{s}_1, \dots, \mathbf{s}_M) \in \mathbf{T}^{(n)}(E).$$

Example 3.4. Let $X_t = t\mathbf{x} \in \mathbb{R}^d$, then

$$\mathbf{Sig}_{[0,1]}(X) = (1, \mathbf{x}, \frac{\mathbf{x}^{\otimes 2}}{2!}, \dots) \in \mathbf{T}(E)$$

3.1.1 Signature map

We expect a good feature map to capture important information while ignoring irrelevant ones. First observation is the the signature feature is invariant of starting point because $d(X_t - X_0) = dX_t$. Moreover, it is invariant of reparametrization.

Proposition 3.5 (Invariant under reparametrization). *Let $X \in \mathcal{C}_0^1([S_1, T_1], E)$ and $\tau: [S_1, T_1] \rightarrow [S_2, T_2]$ a non-decreasing surjective reparametrization. Then*

$$\mathbf{Sig}_{[S_2, T_2]}(X_{\tau(\cdot)}) = \mathbf{Sig}_{[S_1, T_1]}(X) \quad (3.1)$$

Proof.

$$\begin{aligned} \mathbf{Sig}_{[S_2, T_2]}(X_{\tau(\cdot)})_k &= \int_{\tau(t_1) < \dots < \tau(t_k) \in [S_2, T_2]} dX_{\tau(t_1)} \otimes \dots \otimes dX_{\tau(t_k)} \\ &= \int_{t_1 < \dots < t_k \in [S_1, T_1]} dX_{t_1} \otimes \dots \otimes dX_{t_k} = \mathbf{Sig}_{[S_1, T_1]}(X) \end{aligned}$$

□

From the invariant property of reparametrization, we notice that signature map is not injective. However, signature feature is injective up to tree-like equivalence \sim_t which we detailed in appendix, and obviously time-reparametrization is included in tree-like equivalence. We define $\mathcal{P}_0^1 = \mathcal{C}_0^1([0, T], E) / \sim_t$ the quotient space up to tree-like equivalence endowed with quotient metric.

Theorem 3.6 (Weak uniqueness [1]). *$\mathbf{Sig}_{[0, T]}$ is injective on \mathcal{P}_0^1 .*

Proof. See ???

□

If we add time as an additional strictly increasing coordinate into path ie. $\overline{X}_t = (t, X_t) \in \overline{E} := \mathbb{R} \oplus E$, then $\mathbf{Sig}_{[0, T]}$ is injective on the time-augmented space.

Corollary 3.7 (Uniqueness [1]). *$\mathbf{Sig}_{[0, T]}$ is injective on $\mathcal{C}_0^1([0, T], \overline{E})$.*

Proof. Since the first coordinate is strictly increasing, the only tree-like equivalent path is itself i.e. $\mathcal{P}_0^1 = \mathcal{C}_0^1([0, T], \overline{E})$, which concludes the proof. □

Proposition 3.8. *$\mathbf{Sig}_{[0, 1]}$ is neither surjective nor the range of which a linear subspace of $\mathbf{T}(E)$.*

Proof. Let $X \in \mathcal{C}_0^1([0, T], E)$ and w.l.o.g assume $X_0 = 0$. By integration by part

$$\begin{aligned} \mathbf{Sig}_{[0,1]}(X)_{1,2} + \mathbf{Sig}_{[0,1]}(X)_{2,1} &= \int_{t_1 < t_2 \in [0,1]} dX_{t_1}^1 dX_{t_2}^2 + \int_{t_1 < t_2 \in [0,1]} dX_{t_1}^2 dX_{t_2}^1 \\ &= \int_{t \in [0,1]} d(X_t^1 X_t^2) = \mathbf{Sig}_{[0,1]}(X)_1 \mathbf{Sig}_{[0,1]}(X)_2. \end{aligned}$$

Thus, $\mathbf{Sig}_{[0,1]}$ is neither surjective nor the range of it a linear subspace of $\mathbf{T}(E)$. \square

Theorem 3.9 (Weak universality). *Let A be a compact set of $\mathbf{Sig}(\mathcal{C}_0^1(J, E))$, then for all $f: A \rightarrow \mathbb{R}$ continuous and for all $\epsilon > 0$, there exists a linear functional $L \in \mathbf{T}(E)^*$ such that*

$$\sup_{\mathbf{a} \in A} \|f(\mathbf{a}) - L(\mathbf{a})\| \leq \epsilon \quad (3.2)$$

Proof. Proof relies on the Stone-Weierstrass theorem and the shuffle product property of the signature detailed in appendix, see [12]. \square

3.1.2 Signature stream

Instead of viewing signature as a static object, we can consider signature stream of a path $X \in \mathcal{C}_0^1([0, T], E)$ as a process $(\mathbf{Sig}_{[0,t]}(X))_{[0,T]}$ taking values in $\mathbf{T}(E)$.

Proposition 3.10. *Let $X \in \mathcal{C}_0^1([0, T], E)$ and define $\pi_n: \mathbf{T}(E) \rightarrow \mathbf{T}^{(n)}(E)$ the projection such that for all $\mathbf{x} \in \mathbf{T}(E)$*

$$\pi_n((\mathbf{x}_k)_{k \geq 0}) = (\mathbf{x}_k)_{k \leq n} \quad (3.3)$$

then $S_t = \mathbf{Sig}_{[0,t]}^{(n)}(X)$ satisfies for all $t \in [0, T]$ that

$$dS_t = \pi_n(S_t \otimes dX_t), \quad S_0 = (1, 0, \dots), \quad (3.4)$$

and moreover $(S_t)_{t \in [0, T]}$ is the unique solution of (3.4).

Proof. See Lemma 2.10 in [15]. \square

Definition 3.11. Let $X \in \mathcal{C}_0^1([0, s], E)$ and $Y \in \mathcal{C}_0^1([s, t], E)$. The concatenated path $X \star Y \in \mathcal{C}_0^1([0, t], E)$ is defined by

$$(X \star Y)_u = \begin{cases} X_u & u \in [0, s] \\ Y_u + (X_s - Y_s) & u \in [s, t] \end{cases} \quad (3.5)$$

Theorem 3.12 (Chen's identity). *Let $X \in \mathcal{C}_0^1([0, s], E)$ and $Y \in \mathcal{C}_0^1([s, t], E)$. Then*

$$\mathbf{Sig}_{[0,t]}(X \star Y) = \mathbf{Sig}_{[0,s]}(X) \otimes \mathbf{Sig}_{[s,t]}(Y) \quad (3.6)$$

Proof. See Theorem 2.9 in [15]. \square

Example 3.13. *Let X be linear on $[n, n+1]$ and let $X_{n+1} - X_n = \mathbf{x}_n$ for $n \in \mathbb{N}$, then*

$$\mathbf{Sig}_{[0,N]}(X) = \bigotimes_{n \leq N} (1, \mathbf{x}_n, \frac{\mathbf{x}_n^{\otimes 2}}{2!}, \dots)$$

Example 3.14 (Linear controlled differential equation). *Let $E = \mathbb{R}^d, W = \mathbb{R}^n$. let $X \in \mathcal{C}_0^1([0, T], E)$ and let $B: E \rightarrow \mathbf{L}(W)$ be a bounded linear map. Consider*

$$dY_t = B(dX_t)(Y_t), \quad Y_0 \in W \quad (3.7)$$

If we denote $B^k := B(e_k)$, $k = 1, \dots, d$ then

$$dY_t = \sum_{k=1}^d B^k(Y_t) dX_t^k, \quad Y_0 \in W. \quad (3.8)$$

It follows from Picard's iteration that

$$\begin{aligned} Y_t^n &= \left(I + \sum_{k=1}^n B^{\otimes k} \int_{t_1 < \dots < t_k \in [0,t]} dX_{t_1} \otimes \dots \otimes dX_{t_k} \right) Y_0 \\ &= \left(I + \sum_{k=1}^n \sum_{i_1, \dots, i_k=1}^d B^{i_k} \dots B^{i_1} \int_{t_1 < \dots < t_k \in [0,t]} dX_{t_1}^{i_1} \dots dX_{t_k}^{i_k} \right) Y_0. \end{aligned} \quad (3.9)$$

Let the variation of $X \in \mathcal{C}_0^1([0, T], E)$ denoted by $\|X\|_{[0,T]}$, then

$$\left\| \int_{t_1 < \dots < t_k \in [0,t]} dX_{t_1} \otimes \dots \otimes dX_{t_k} \right\|_{E^{\otimes k}} \leq \frac{\|X\|_{[0,T]}^k}{k!}. \quad (3.10)$$

Therefore, Y_t^n converges to Y_t as $n \rightarrow \infty$ i.e.

$$\|Y_t - Y_t^n\|_W \leq \sum_{k>n} \frac{\|B\|_{\mathcal{L}(E, \mathcal{L}(W))}^k \|X\|_{[0,T]}^k}{k!} \leq \frac{\|B\|_{\mathcal{L}(E, \mathcal{L}(W))}^{n+1} \|X\|_{[0,T]}^{n+1}}{n!} \rightarrow 0, \quad \text{as } n \rightarrow \infty \quad (3.11)$$

and

$$Y_t = \left(I + \sum_{k=1}^{\infty} B^{\otimes k} \int_{t_1 < \dots < t_k \in [0, t]} dX_{t_1} \otimes \dots \otimes dX_{t_k} \right) Y_0. \quad (3.12)$$

In the language of signature

$$Y_t = \left(\sum_{k=0}^{\infty} B^{\otimes k} \right) \left(\mathbf{Sig}_{[0, t]}(X) \right) Y_0. \quad (3.13)$$

which implies that the solution of controlled SDE could be written as a linear function on signature stream of control path. This implies that signature stream is a promising feature for controlled ODE.

Remark. Similar result holds for smooth vector field with some additional boundedness assumption and the proof is in alignment with the argument above, see [12].

Despite nice properties as feature map, signature feature suffers from dimension explosion w.r.t truncation order. Consider a path $X \in \mathcal{C}_0^1([0, T], \mathbb{R}^d)$ and the truncated signature of X up to order n in $\mathbf{T}^{(n)}(\mathbb{R}^d) \subseteq \bigoplus_{k \leq n} (\mathbb{R}^d)^{\otimes k}$ which has dimension $1 + d + \dots + d^n = \frac{d^{n+1} - d}{d - 1}$ growing exponentially w.r.t the truncation order n . Therefore, we would like to explore some low-rank approximations of signature feature.

3.2 Log-signature

From Proposition 3.8, we know that $\mathbf{Sig}_{[0, 1]}$ is neither surjective nor the range of which a linear subspace of $\mathbf{T}(E)$. In Example 3.4, for $X_t = t\mathbf{x} \in \mathbb{R}^d$, then

$$\mathbf{Sig}_{[0, 1]}(X) = \sum_{k \geq 0} \frac{\mathbf{x}^{\otimes k}}{k!}$$

which is the power series expansion of exponential function. Motivated by this, we curve the signature space by taking ‘logarithm’ which we defined below.

Definition 3.15. Define $\mathbf{T}_1(E) = \{\mathbf{a} \in \mathbf{T}(E) : \mathbf{a}_0 = 1\}$. Let $\mathbf{a} \in \mathbf{T}_1(E)$, we define the exponential be

$$\exp(\mathbf{a}) = \sum_{n \geq 0} \frac{\mathbf{a}^{\otimes n}}{n!}.$$

and we define the logarithm be

$$\log(\mathbf{a}) = \log(1 + \mathbf{t}) = \sum_{n \geq 1} \frac{(-1)^{n-1}}{n} \mathbf{t}^{\otimes n}.$$

Lemma 3.16. $\exp(\cdot)$ and $\log(\cdot)$ are inverse of each other on $\mathbf{T}_1(E)$.

Proof. See Lemma 2.21 in [15]. □

Therefore, taking logarithm of signature loses no information.

Definition 3.17 (Log-signature). We denote $\mathbf{LogSig}_J: \mathcal{C}_0^1(J, E) \rightarrow \mathbf{T}(E)$ the log-signature map such that for all $X \in \mathcal{C}_0^1(J, E)$

$$\mathbf{LogSig}_J(X) = \log(\mathbf{Sig}_J(X))$$

Let $\mathbf{LogSig}_J^{(n)}$ denote the truncated signature map up to order n

$$\mathbf{LogSig}_J^{(n)}(X) = \pi_n(\mathbf{LogSig}_J(X))$$

where π_n define in (3.3).

3.2.1 Lie Series

Besides storing the same information, thanks to the shuffle product property of signature, log-signature space is linear and admits a more concrete representation. Precisely speaking, log-signature space is a linear subspace of Lie formal series over E . Recall the Lie bracket $[\cdot, \cdot]$ on $T((E))$ such that

$$[\mathbf{a}, \mathbf{b}] = \mathbf{a} \otimes \mathbf{b} - \mathbf{b} \otimes \mathbf{a}$$

If F_1 and F_2 are two linear subspaces of $T((E))$, let us denote by $[F_1, F_2]$ the linear span of all the elements of the form $[\mathbf{a}, \mathbf{b}]$, where $\mathbf{a} \in F_1$ and $\mathbf{b} \in F_2$.

Definition 3.18 (Lie formal series). We denote $\mathcal{L}((E))$ the space of Lie formal series over E such that

$$\mathcal{L}((E)) = \{\mathbf{l} \in T((E)) : \mathbf{l}_n \in L_n\}$$

where

$$L_0 = 0, L_1 = [E, E], L_2 = [E, L_1], \dots, L_n = [E, L_{n-1}], \dots$$

and denote $\mathcal{L}^{(n)}(E) = \pi_n(\mathcal{L}((E)))$ the truncated Lie series up to order n .

Theorem 3.19. *The range of log-signature is a linear subspace of Lie series over E .*

$$\mathbf{LogSig}(\mathcal{C}_0^1) \subseteq \mathcal{L}((E)).$$

Proof. See Theorem 2.23 in [15]. □

Theorem 3.20. *The range of truncated log-signature is the truncated Lie series over E up to the same order*

$$\mathbf{LogSig}^{(n)}(\mathcal{C}_0^1) = \mathcal{L}^{(n)}(E)$$

Proof. See Proposition 2.27 in [15]. □

With Lie series form, we compute the dimension of truncated log-signature space.

Proposition 3.21. *The dimension of the space of truncated log-signature up to order n is*

$$w(d, n) = \sum_{k=1}^n \frac{1}{k} \sum_{i|k} \mu\left(\frac{k}{i}\right) d^i$$

which is the Witt's formula and μ is the Möbius function.

Proof. See Corollary 4.14 in [16]. □

If we let $d = 5$ and $n = 5$, then $\dim(\mathbf{Sig}) = 3905$ while $\dim(\mathbf{LogSig}) = 829$. In summary, log-signature curves the signature space and reduces the redundancy without losing any information. However, there is no free lunch because log-signature loses the universality of signature, therefore requiring nonlinear models.

Example 3.22. *For $X_t = t\mathbf{x} \in \mathbb{R}^d$, then*

$$\mathbf{Sig}_{[0,1]}(X) = \exp(\mathbf{x}), \quad \mathbf{LogSig}_{[0,1]}(X) = \mathbf{x}.$$

Log-signature pick non-redundant information in signature.

3.3 Universality and characteristics

Recall the weak universality of signature, we restrict ourselves on a compact set of $\mathbf{Sig}(\mathcal{C}_0^1(J, E))$. To generalize the theorem beyond compactness, we would like to normalize the set $\mathbf{Sig}(\mathcal{C}_0^1(J, E))$ into a bounded ball. The first idea comes to mind is to normalize signatures by scaling them on $\mathbf{T}(E)$. However, because signature map is neither surjective nor the range of it a linear subspace, scaled tensor of a signature might not be signature. Thus, we drive ourselves out of the region of $\mathbf{Sig}(\mathcal{C}_0^1(J, E))$ where we have the shuffle product property. An alternative is to normalize the

signature by scaling the path. For $X \in \mathcal{C}_0^1(J, E)$ and $\lambda > 0$, the signature of the scaled path

$$\begin{aligned}\mathbf{Sig}_J(\lambda X)_k &= \int_{t_1 \leq \dots \leq t_k \in J} d\lambda X_{t_1} \otimes \dots \otimes d\lambda X_{t_k} \\ &= \lambda^k \int_{t_1 \leq \dots \leq t_k \in J} dX_{t_1} \otimes \dots \otimes dX_{t_k} = \lambda^k \mathbf{Sig}_J(X)_k\end{aligned}$$

Thus, the norm of the scaled path

$$\|\mathbf{Sig}_J(\lambda X)\|_{\mathbf{T}(E)}^2 = \sum_{k \geq 0} \lambda^{2k} \|\mathbf{Sig}_J(X)_k\|_{E^{\otimes k}}^2$$

Therefore, we define the scaling map δ_λ on $\mathbf{T}(V)$ such that for all $\mathbf{a} \in \mathbf{T}(V)$

$$\delta_\lambda(\mathbf{a}) = (\lambda^k \mathbf{a}^k)_{k \geq 0}$$

Definition 3.23. A tensor normalization is a continuous injective map

$$\Lambda: \mathbf{T}(V) \rightarrow \{\mathbf{t} \in \mathbf{T}(V): \|\mathbf{t}\|_{\mathbf{T}(E)} \leq K\}$$

$$\mathbf{t} \mapsto \delta_{\lambda(\mathbf{t})}(\mathbf{t})$$

where $K > 0$ and $\lambda: \mathbf{T}(V) \rightarrow (0, \infty)$ a function.

The existence of tensor normalization is not trivial because of the nonlinear relationship between the scaling factor $\lambda(\mathbf{t})$ and the norm of scaled tensor $\|\delta_{\lambda(\mathbf{t})}(\mathbf{t})\|$ shown above. The proof of existence and the construction methodology can be found in the appendix.

Theorem 3.24 ([3]). *Let $\Lambda: \mathbf{T}(E) \rightarrow \mathbf{T}(E)$ be a tensor normalization. The normalized signature*

$$\Phi = \Lambda \circ \mathbf{Sig}_J$$

- (i) *is a continuous injection from \mathcal{P}_0^1 in to a bounded subset of $\mathbf{T}(E)$,*
- (ii) *is universal to $C_b(\mathcal{P}_0^1, \mathbb{R})$, equipped with the strict topology,*
- (iii) *is characteristic to the space of finite regular Borel measures on \mathcal{P}_0^1 .*

Proof. See Proposition 4.1 in [3]. □

Corollary 3.25. *Let X a stochastic process on $[0, 1]$ and measurable (Ω, \mathcal{F}) . Let \mathbb{P} and \mathbb{Q} be two regular probability measures such that $X \in \mathcal{P}_0^1$ almost surely. Then*

$$\mathbb{E}_{\mathbb{P}}[\Phi(X)] = \mathbb{E}_{\mathbb{Q}}[\Phi(X)] \quad \text{iff} \quad \mathbb{P} = \mathbb{Q}$$

Proof. Proof directly from (iii) of Theorem 3.24. \square

Remark. If we consider the regular probability measure on time augmented path space $\mathcal{C}_0^1([0, T], \overline{E})$, similar result holds since $\mathcal{C}_0^1([0, T], \overline{E}) = \mathcal{P}_0^1$.

Definition 3.26 (Maximum mean distance). We define the maximum mean distance (MMD) as

$$d_{\mathcal{G}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{G}} |\mathbb{E}_{\mathbb{P}}[f] - \mathbb{E}_{\mathbb{Q}}[f]|$$

where $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{P}_0^1}$.

Let $\mathcal{G} = \{\langle \mathbf{l}, \Phi(\cdot) \rangle_{\mathbf{T}_1} : \mathbf{l} \in \mathbf{T}_1, \|\mathbf{l}\|_{\mathbf{T}_1} \leq 1\}$, then by Corollary 3.25

$$d_{\mathcal{G}}(\mathbb{P}, \mathbb{Q}) = 0 \quad \text{iff} \quad \mathbb{P} = \mathbb{Q}$$

which implies that a metric, see [5]. Moreover

$$\begin{aligned} d_{\mathcal{G}}(\mathbb{P}, \mathbb{Q}) &= \sup_{f \in \mathcal{G}} |\mathbb{E}_{\mathbb{P}}[f] - \mathbb{E}_{\mathbb{Q}}[f]| \\ &= \sup_{f \in \mathcal{G}} \left| \mathbb{E}_{(X,Y) \sim \mathbb{P} \otimes \mathbb{Q}} [f(X) - f(Y)] \right| \\ &= \sup_{\mathbf{l} \in \mathbf{T}_1, \|\mathbf{l}\|_{\mathbf{T}_1} \leq 1} \left| \mathbb{E}_{(X,Y) \sim \mathbb{P} \otimes \mathbb{Q}} [\langle \mathbf{l}, \Phi(X) - \Phi(Y) \rangle_{\mathbf{T}_1}] \right| \end{aligned}$$

Since $\mathbb{E}_{(X,Y) \sim \mathbb{P} \otimes \mathbb{Q}} [\Phi(X) - \Phi(Y)] \in \mathbf{T}_1$, then

$$d_{\mathcal{G}}(\mathbb{P}, \mathbb{Q}) = \mathbb{E}[\langle \Phi(X) - \Phi(Y), \Phi(X') - \Phi(Y') \rangle_{\mathbf{T}_1}]$$

where X, Y, X', Y' are independent with $X, X' \sim \mathbb{P}$ and $Y, Y' \sim \mathbb{Q}$

Definition 3.27. We denote $\mathbf{k}_{\text{Sig}} : \mathbf{T}_1 \times \mathbf{T}_1 \rightarrow \mathbb{R}$ the signature kernel such that

$$\mathbf{k}_{\text{Sig}}(\cdot, \cdot) = \langle \Phi(\cdot), \Phi(\cdot) \rangle_{\mathbf{T}_1}$$

Then we can rewrite the MMD as

$$d_{\mathcal{G}}(\mathbb{P}, \mathbb{Q}) = \mathbb{E}[\mathbf{k}_{\text{Sig}}(X, X')] - 2\mathbb{E}[\mathbf{k}_{\text{Sig}}(X, Y)] + \mathbb{E}[\mathbf{k}_{\text{Sig}}(Y, Y')]$$

where X, Y, X', Y' are independent with $X, X' \sim \mathbb{P}$ and $Y, Y' \sim \mathbb{Q}$. Similar to the kernel trick in machine learning, the kernel representation of MMD provides a very efficient way evaluating the metric, which we will elaborate in the next section.

3.4 Discrete signature feature

Let $X \in \mathcal{C}_0^1([0, 1], E)$ and consider a discrete sequence $\mathbf{X} = (X_0, X_{1/N}, \dots, X_1) \in E^{N+1}$. We extend the concept of signature on discrete sequence E^{N+1} by computing signature of the linear interpolation of sequence. From Example 3.13 we know the signature of piecewise-linear path is

$$\exp(\mathbf{X}_1 - \mathbf{X}_0) \otimes \dots \otimes \exp(\mathbf{X}_N - \mathbf{X}_{N-1})$$

Therefore we define the discrete signature as follows

Definition 3.28. Let $\mathbf{X} = (\mathbf{X}_i)_{i=0}^N \in E^{N+1}$ and let $\Delta \mathbf{X}_i = \mathbf{X}_i - \mathbf{X}_{i-1}$. We denote $\mathbf{sig}_{[0, N]}$ the discrete signature of order m and depth n such that

$$\mathbf{sig}_{[0, N]}(\mathbf{X}) = \pi_n \left(\bigotimes_{i=1}^N \pi_m \left(\exp(\Delta \mathbf{X}_i) \right) \right)$$

In particular, if $m = 1$ we call it growing discrete signature

$$\begin{aligned} \mathbf{sig}_{[0, N]}(\mathbf{X}) &= \pi_n \left(\prod_{i=1}^N (1 + \Delta \mathbf{X}_i) \right) \\ &= \sum_{k=0}^n \sum_{i_1 < \dots < i_k=1}^N \Delta \mathbf{X}_{i_1} \otimes \dots \otimes \Delta \mathbf{X}_{i_k}. \end{aligned}$$

If $m > 1$, we call it high order discrete signature, and if $m \geq n$, we call it truncated discrete signature

$$\begin{aligned} \mathbf{sig}_{[0, N]}(\mathbf{X}) &= \pi_n \left(\prod_{i=1}^N \exp(\Delta \mathbf{X}_i) \right) \\ &= \sum_{k=0}^n \sum_{i_1 \leq \dots \leq i_k=1}^N \Delta \mathbf{X}_{i_1} \otimes \dots \otimes \Delta \mathbf{X}_{i_k} \end{aligned}$$

Lemma 3.29. Let $X \in \mathcal{C}_0^1([0, 1], E)$ and $\mathbf{X} = (X_{i/N})_{i=0}^N \in E^{N+1}$, then

$$\|\mathbf{Sig}_{[0, 1]}(X) - \mathbf{sig}_{[0, N]}(\mathbf{X})\|_{\mathbf{T}_1} \rightarrow 0 \quad \text{as } N \rightarrow \infty$$

Proof. See Theorem 5 in [10]. □

Remark. If we let X to be the Brownian motion (not in \mathcal{C}_0^1), the growing discrete signature converges to the signature defined with Ito integral, while the high order discrete signature converges to the signature defined with Stratonovich integral. More generally, if we want to approximate a geometric p -rough path, we need high order discrete signature at least order $\lceil p \rceil$, see [10].

3.4.1 Kernel trick

In the same fashion of horner algorithm

$$\begin{aligned} & a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots + a_nx^n \\ &= a_0 + x \left(a_1 + x \left(a_2 + x \left(a_3 + \cdots + x(a_{n-1} + x a_n) \cdots \right) \right) \right). \end{aligned}$$

we can reduce the computational complexity of discrete signature by dynamic programming principle.

Proposition 3.30. *Let $\mathbf{X} = (\mathbf{X}_i)_{i=0}^N \in E^{N+1}$ and let $\mathbf{sig}_{[0,N]}$ be the growing signature*

$$\begin{aligned} \mathbf{sig}_{[0,N]}(\mathbf{X}) &= \mathbf{sig}_{[0,N-1]}(\mathbf{X}) \otimes (1 + \Delta \mathbf{X}_N) \\ &= 1 + \sum_{i_1=1}^N \Delta \mathbf{X}_{i_1} \left(1 + \sum_{i_2=i_1+1}^N \Delta \mathbf{X}_{i_2} (\cdots) \right) \end{aligned}$$

This is too surprise because we just rewrite the Chen's identity. However, this greatly reduce the complexity when computing the kernel.

Definition 3.31. Let $\mathbf{sig}_{[0,N]}$ be the discrete signature, then we denote $\mathbf{k}_{\mathbf{sig}}$ the discrete signature kernel such that

$$\mathbf{k}_{\mathbf{sig}}(\cdot, \cdot) = \langle \mathbf{sig}_{[0,N]}(\cdot), \mathbf{sig}_{[0,N]}(\cdot) \rangle_{\mathbf{T}_1}$$

Proposition 3.32. *Let $\mathbf{X} = (\mathbf{X}_i)_{i=0}^N \in E^{N+1}$ and let $\mathbf{Y} = (\mathbf{Y}_i)_{i=0}^N \in E^{N+1}$. Let $\mathbf{k}_{\mathbf{sig}}$ be the kernel of discrete growing signature*

$$\mathbf{k}_{\mathbf{sig}}(\mathbf{X}, \mathbf{Y}) = 1 + \sum_{\substack{i_1=1 \\ j_1=1}}^N \langle \Delta \mathbf{X}_{i_1}, \Delta \mathbf{Y}_{j_1} \rangle \left(1 + \sum_{\substack{i_2=i_1+1 \\ j_2=j_1+1}}^N \langle \Delta \mathbf{X}_{i_2}, \Delta \mathbf{Y}_{j_2} \rangle (\cdots) \right) \quad (3.14)$$

Remark. Similar formulas hold for high order discrete signature and kernel, see [10].

This result is important for numerical implementation in two aspects. First is that we avoid computing the tensor but the inner product directly by kernel trick. If somehow (for better characteristic capacity) you would like lift paths to a RKHS space (H, κ) before applying the signature kernel, this formula avoids explicitly compute $\Delta \kappa_X$ and $\Delta \kappa_Y$. but approximating with

$$\langle \Delta \kappa_{X_i}, \Delta \kappa_{Y_j} \rangle \approx \kappa(X_{i+1}, Y_{j+1}) + \kappa(X_i, Y_j) - \kappa(X_i, Y_{j+1}) - \kappa(X_{i+1}, Y_j),$$

which reduce a potentially infinite computational complexity if (H, κ) is an infinite dimensional space, see [3]. Secondly, the recursive structure in time implies that computing the inner product of signature stream is as cheap as computing the inner product of signature. This recursive feature can also be used for more general inner product between tensor and discrete signature.

3.5 Low-rank-signature

We extend the recursive method from discrete signature to tensor with similar recursive structure. Let $\mathbf{l} = (\mathbf{l}_k)_{k=0}^\infty \in \mathbf{T}_1(E)$, let $\mathbf{X} = (\mathbf{X}_i)_{i=0}^N \in E^{N+1}$, and let $\mathbf{sig}_{[0,N]}$ be the growing signature. If $\mathbf{l}_k = \mathbf{l}_{k-1} \otimes l_k$ then

$$\begin{aligned} \langle \mathbf{l}, \mathbf{sig}_{[0,N]}(\mathbf{X}) \rangle_{\mathbf{T}_1(E)} &= \sum_{k \geq 0} \langle \mathbf{l}_k, \mathbf{sig}_{[0,N]}^{(k)}(\mathbf{X}) \rangle_{\mathbf{T}_1(E)} \\ &= \sum_{k \geq 0} \sum_{i=1}^N \langle \mathbf{l}_{k-1}, \mathbf{sig}_{[0,i-1]}^{(k-1)}(\mathbf{X}) \rangle_{\mathbf{T}_1(E)} \cdot \langle l_k, \Delta \mathbf{X}_i \rangle_{\mathbf{T}_1(E)} \\ &= 1 + \sum_{\substack{i_1=1 \\ j_1=1}}^N \langle l_{i_1}, \Delta \mathbf{X}_{j_1} \rangle \left(1 + \sum_{\substack{i_2=i_1+1 \\ j_2=j_1+1}}^N \langle l_{i_2}, \Delta \mathbf{X}_{j_2} \rangle (\dots) \right) \end{aligned}$$

This is exactly the formula (3.14) in Proposition 3.32. In the intermediate step of computing the inner product, we have obtained the value of $\langle \mathbf{l}_k, \mathbf{sig}_{[0,N]}^{(k)}(\mathbf{X}) \rangle_{\mathbf{T}_1(E)}$ for all k . Therefore, if we consider a sequence of L many tensors such that $\mathbf{l}_k = \mathbf{l}_{k-1} \otimes l_k$, then we can leverage the recursive structure in order and compute their inner product with signature in linear computational complexity w.r.t L .

Remark. This result can also be generalized to higher order case by considering the recursive algorithm computing high order discrete signature and kernel, see [10]. However, in the paper [17] introducing low-rank tensor projection of ordered sequential data, only first order recursive structure is considered. We believe that higher order recursive structure also exists and we wish to prove this claim in further research.

If we only assume $\mathbf{l}_k = l_{k,1} \otimes \dots \otimes l_{k,k}$, then we lose the recursive in order, but still we have the recursive in time of discrete signature and the kernel trick.

$$\langle \mathbf{l}_k, \mathbf{sig}_{[0,N]}^{(k)}(\mathbf{X}) \rangle_{\mathbf{T}_1(E)} = \sum_{i=1}^N \langle l_{k,1} \otimes \dots \otimes l_{k-1,k}, \mathbf{sig}_{[0,i-1]}^{(k-1)}(\mathbf{X}) \rangle_{\mathbf{T}_1(E)} \cdot \langle l_{k,k}, \Delta \mathbf{X}_i \rangle_{\mathbf{T}_1(E)}$$

Definition 3.33. We call $\mathbf{l} \in \mathbf{T}_1$ rank-1 tensor if

$$\mathbf{l} = l_1 \otimes \cdots \otimes l_k, \quad \text{for some } k \geq 1$$

Similar to log-signature, the inner product between truncated signature and rank-1 tensor map the signature space to a low dimensional space \mathbb{R} . This projection is linear, computational cheap, and more importantly trainable. The choice of rank-1 tensors gives us the freedom to choose the output space dimension as well as the direction of projection.

Definition 3.34. Let $(\mathbf{l}_l)_{l=0}^L$ a sequence of rank-1 tensors in \mathbf{T}_1 and let $\mathbf{sig}_{[0,N]}$ the discrete signature, then we denote $\mathbf{LRsig}_{[0,N]}$ the low-rank signature such that

$$\mathbf{LRsig}_{[0,N]} = (\langle \mathbf{l}_l, \mathbf{sig}_{[0,N]} \rangle_{\mathbf{T}_1(E)})_{l=0}^L \in \mathbb{R}^L$$

and we define the low-rank signature stream $(\mathbf{LRsig}_{[0,i]})_{i=0}^N$.

From the recursive algorithm above, we know that computing the low-rank signature stream is as cheap as computing the low-rank signature. Unlike signature stream, low-rank signature stream still in a low dimensional space, so we may again apply the low-rank signature stream to it. Thus, we may stack many levels of signature stream on the original path and this has been proved successful dealing with sequential data, see [17].

4. Variational Autoencoders

In this chapter, we give a brief introduction to the variational autoencoders (VAEs) and some variations of VAE such as conditional VAEs (CVAEs), β -VAEs, student-VAEs, and gaussian process VAEs (GP-VAEs). Variational autoencoders is a generative model which simulates how the data is generated under certain presumed distribution. In short, VAEs are variational bayesian inference on latent variable models with likelihood and posterior described by neural networks. In the scenario of generating time series, there are in general two frameworks to apply variational autoencoders. First is to assume a bayesian model with both observable variables and latent variables taking values in space of time series, for example taking latent space as a gaussian process in GP-VAEs. Another approach is to first consider an appropriate feature map from time series to a finite dimensional features space and apply VAE on this feature space. After the training of VAE on features space, we generate new samples on the feature space and transform those samples on feature space back to time series with the inverse map of the feature map. More details and background information on VAEs can be found in

4.1 Variational autoencoders

Variational autoencoders are variational inference on deep latent variable model. Let consider deep latent variable models consists of observable random variable x , latent random variable z , and parameter θ . The joint distribution $p_\theta(x, z)$ of the deep latent variable model is parametrized by neural networks. For example, we may choose z be standard normal distribution and x a normal distribution with mean $\mu = \mathbf{NN}(z)$ and variance 1, where \mathbf{NN} is a neural network. Moreover, we call $p_\theta(z) = p(z)$ the *prior* (independent with θ), $p_\theta(z|x)$ the *posterior*, $p_\theta(x|z)$ the *conditional distribution* and $p_\theta(x)$ the *marginal likelihood*. The big advantage of deep latent variable model

is that even when the prior $p(z)$ and conditional distribution $p(x|z)$ are explicit and simple, the marginal likelihood $p_\theta(x)$ can be very expressive due to the universal approximating capacity of neural network. However, the price to pay for expressive marginal likelihood is the intractability of marginal likelihood, namely having no analytic solution or efficient estimator for it. By Bayes' rule

$$p_\theta(x) = \frac{p_\theta(x, z)}{p_\theta(z|x)}.$$

The joint distribution is not hard to write down since

$$p_\theta(x, z) = p(x|\mathbf{NN}_\theta(z))p(z)$$

but the posterior $p_\theta(z|x)$ in general has no analytic solution or efficient estimator. Therefore, we need to leverage the idea of variational inference, introducing a distribution $q_\phi(z|x)$ approximating the true posterior $p_\theta(z|x)$.

4.1.1 Variational inference

Let x_1, \dots, x_n be i.i.d. observable samples drawn from a random variable x . Maximizing the log-marginal likelihood by Bayes's rule, we obtain

$$\begin{aligned} \log p_\theta(x_i) &= \log \left(\frac{p_\theta(x_i, z)}{p_\theta(z|x_i)} \right) \\ &= \log \left(\frac{p_\theta(x_i, z)}{q_\phi(z|x_i)} \frac{q_\phi(z|x_i)}{p_\theta(z|x_i)} \right) \\ &= \mathbb{E}_{q_\phi(z|x_i)} \left[\log \left(\frac{p_\theta(x_i, z)}{q_\phi(z|x_i)} \right) + \log \left(\frac{q_\phi(z|x_i)}{p_\theta(z|x_i)} \right) \right] \\ &= \mathbb{E}_{q_\phi(z|x_i)} \left[\log \left(\frac{p_\theta(x_i, z)}{q_\phi(z|x_i)} \right) + \log \left(\frac{q_\phi(z|x_i)}{p_\theta(z|x_i)} \right) \right] \\ &= \mathbb{E}_{q_\phi(z|x_i)} \left[\log \left(\frac{p_\theta(x_i, z)}{q_\phi(z|x_i)} \right) \right] + KL(q_\phi(z|x_i) || p_\theta(z|x_i)) \\ &\geq \mathbb{E}_{q_\phi(z|x_i)} \left[\log \left(\frac{p_\theta(x_i, z)}{q_\phi(z|x_i)} \right) \right] \end{aligned}$$

Let us denote that $\mathcal{L}_{\theta, \phi}(x)$ the *evidence lower bound* (ELBO) such that

$$\mathcal{L}_{\theta, \phi}(x) = \mathbb{E}_{q_\phi(z|x)} \left[\log \left(\frac{p_\theta(x, z)}{q_\phi(z|x)} \right) \right]$$

Notice that

$$\log p_\theta(x) = \max_{\phi} \mathcal{L}_{\theta, \phi}(x).$$

Thus, we can solve the maximization problem of ELBO instead of the marginal likelihood. Also if

$$q_\phi^*(z|x) \in \arg \max \mathcal{L}_{\theta, \phi}(x)$$

then it satisfies that $q_\phi^*(z|x) = p_\theta(z|x)$ because

$$KL(q_\phi(z|x_i) || p_\theta(z|x_i)) = 0 \quad \text{iff} \quad q_\phi(z|x) = p_\theta(z|x).$$

Therefore maximizing the ELBO also gives us the true posterior. Moreover, observe that

$$\begin{aligned} \mathcal{L}_{\theta, \phi}(x) &= \mathbb{E}_{q_\phi(z|x)} \left[\log \left(\frac{p_\theta(x, z)}{q_\phi(z|x)} \right) \right] \\ &= \mathbb{E}_{q_\phi(z|x)} \left[\log \left(\frac{p_\theta(x|z)p(z)}{q_\phi(z|x)} \right) \right] \\ &= -KL(q_\phi(z|x) || p_\theta(z)) + \mathbb{E}_{q_\phi(z|x)} [p_\theta(x|z)]. \end{aligned}$$

Thus, we are minimizing the KL-distance between approximating posterior and true prior, meanwhile maximizing expected conditional distribution under approximating posterior distribution.

4.1.2 Reparameterization Trick

In order to solve the variational problem, we compute the gradient of the objective function w.r.t θ and ϕ . However, the gradient w.r.t ϕ is in general hard to compute because the expectation also depends on ϕ .

$$\begin{aligned} \nabla_{\phi} \mathcal{L}_{\theta, \phi}(x) &= \nabla_{\phi} \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x, z) - \log q_\phi(z|x)] \\ &\neq \mathbb{E}_{q_\phi(z|x)} [\nabla_{\phi} (\log p_\theta(x, z) - \log q_\phi(z|x))] \end{aligned}$$

Therefore, we need to factorize the randomness taking expectation out of the parameter ϕ . So we consider

$$z = g(\epsilon, \phi, x)$$

where g is a differentiable function and ϵ is independent with ϕ and x . In this case, for all differentiable function f

$$\begin{aligned}
\nabla_{\phi} \mathcal{L}_{\theta, \phi}(x) &= \nabla_{\phi} \mathbb{E}_{q_{\phi}(z|x)} \left[f(x, z) \right] \\
&= \nabla_{\phi} \mathbb{E}_{g(\epsilon, \phi, x)} \left[f(x, z) \right] \\
&= \nabla_{\phi} \mathbb{E}_{\epsilon} \left[f(x, g(\epsilon, \phi, x)) \right] \\
&= \mathbb{E}_{\epsilon} \left[\nabla_{\phi} f(x, g(\epsilon, \phi, x)) \right]
\end{aligned}$$

Example 4.1. Consider a deep latent variable model with gaussian prior and gaussian conditional distribution:

$$\begin{aligned}
q_{\phi}(z|x) &= \mathcal{N}\left(z; \mu_x, \sigma_x^2\right) \\
p_{\theta}(z) &= \mathcal{N}\left(z; 0, I_d\right)
\end{aligned}$$

where $\mu_x \in \mathbb{R}^d$ and σ_x^2 be a diagonal matrix in $\mathbb{R}^{d \times d}$. Then

$$\begin{aligned}
- KL(q_{\phi}(z|x) || p_{\theta}(z)) &= \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(z)] - \mathbb{E}_{q_{\phi}(z|x)} [\log q_{\phi}(z|x)] \\
&= \int \log p_{\theta}(z) q_{\phi}(z|x) dz - \int \log p_{\theta}(z|x) q_{\phi}(z|x) dz \\
&= \frac{1}{2} \sum_{i=1}^d \left(1 + \log(\sigma_i^2) - \mu_i^2 - \sigma_i^2 \right)
\end{aligned}$$

4.2 Variations of VAE

5. Main method

5.1 Time series processing

5.2 Generative model with heavy tail

5.3 Inverse the log-signature

5.4 Evaluation with lifting and normalizing

6. Signature stimulation

6.1 Controlled stochastic differential equation

6.2 Numerical Implementation

7. Conclusions and Future Work

8. Appendix

Bibliography

- [1] BOEDIHARDJO, H., GENG, X., LYONS, T., AND YANG, D. The signature of a rough path: uniqueness. *Advances in Mathematics* 293 (2016), 720–737.
- [2] BONNIER, P., LIU, C., AND OBERHAUSER, H. Adapted topologies and higher rank signatures. *arXiv preprint arXiv:2005.08897* (2020).
- [3] CHEVYREV, I., AND OBERHAUSER, H. Signature moments to characterize laws of stochastic processes. *arXiv preprint arXiv:1810.10971* (2018).
- [4] FOSTER, J., LYONS, T., AND OBERHAUSER, H. An optimal polynomial approximation of brownian motion. *SIAM Journal on Numerical Analysis* 58, 3 (2020), 1393–1421.
- [5] FREEMAN, J. Probability metrics and the stability of stochastic models. *Journal of the Operational Research Society* 43, 9 (1993), 923–923.
- [6] FRIZ, P. K., AND HAIRER, M. *A course on rough paths*. Springer, 2020.
- [7] KIDGER, P., FOSTER, J., LI, X., OBERHAUSER, H., AND LYONS, T. Neural sdes as infinite-dimensional gans. *arXiv preprint arXiv:2102.03657* (2021).
- [8] KIDGER, P., AND LYONS, T. Signatory: differentiable computations of the signature and logsignature transforms, on both cpu and gpu. *arXiv preprint arXiv:2001.00706* (2020).
- [9] KIDGER, P., MORRILL, J., FOSTER, J., AND LYONS, T. Neural controlled differential equations for irregular time series. *arXiv preprint arXiv:2005.08926* (2020).

- [10] KIRÁLY, F. J., AND OBERHAUSER, H. Kernels for sequentially ordered data. *Journal of Machine Learning Research* 20, 31 (2019), 1–45.
- [11] LEVIN, D., LYONS, T., AND NI, H. Learning from the past, predicting the statistics for the future, learning an evolving system. *arXiv preprint arXiv:1309.0260* (2013).
- [12] LIAO, S., LYONS, T., YANG, W., AND NI, H. Learning stochastic differential equations using rnn with log signature features. *arXiv preprint arXiv:1908.08286* (2019).
- [13] LYONS, T. Rough paths, signatures and the modelling of functions on streams. *arXiv preprint arXiv:1405.4537* (2014).
- [14] LYONS, T. J. Differential equations driven by rough signals. *Revista Matemática Iberoamericana* 14, 2 (1998), 215–310.
- [15] LYONS, T. J., CARUANA, M., AND LÉVY, T. *Differential equations driven by rough paths*. Springer, 2007.
- [16] REUTENAUER, C. Free lie algebras. In *Handbook of algebra*, vol. 3. Elsevier, 2003, pp. 887–903.
- [17] TOTH, C., BONNIER, P., AND OBERHAUSER, H. Seq2tens: An efficient representation of sequences by low-rank tensor projections. *arXiv preprint arXiv:2006.07027* (2020).
- [18] TOTH, C., AND OBERHAUSER, H. Variational gaussian processes with signature covariances. *arXiv preprint arXiv:1906.08215* (2019).
- [19] TOTH, C., AND OBERHAUSER, H. Bayesian learning from sequential data using gaussian processes with signature covariances. In *International Conference on Machine Learning* (2020), PMLR, pp. 9548–9560.