NEURAL INFORMATION PROCESSING SYSTEMS

Probabilistic Signature Inversion: Learning Conditional Distributions from Truncated Signatures

Junoh Kang[1] Kiseop Lee[2] Bohyung Han[1,3]

[1]ECE & [3]IPAI, Seoul National University [2]Department of Statistics, Purdue University

junoh.kang@snu.ac.kr





Introduction

Signature: a feature map for paths

Definition 1 (Signature)

$$S(X)_{0,T} = \left(S(X)^I_{0,T}\right)_{I \in \mathcal{W}}, \qquad S(X)^I_{0,T} := \int_{0<t_1<\cdots<t_k<T} dX^{i_1}_{t_1} \cdots dX^{i_k}_{t_k},$$

for a path $X\colon [0,T] \to \mathbb{R}^d$ and multi-indices $I = (i_1, \ldots, i_k) \in \mathcal{W} = \bigcup_{k \geq 0} \{1, \ldots, d\}^k$.

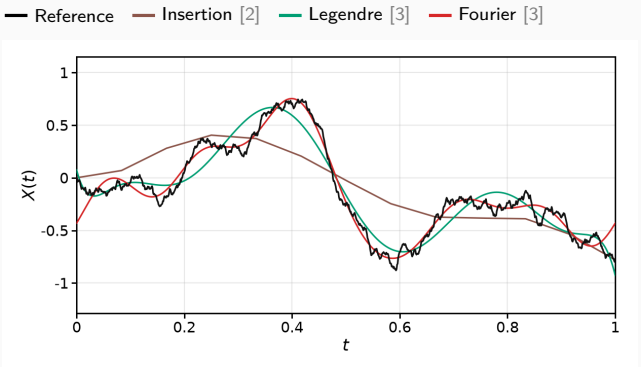
- **Unified representation**
  maps time series of arbitrary lengths or sampling rates to the same space
- **Universality** (linear-friendly)
  $f(X) \approx w^\top S(X)$ for any continuous f
- **Uniqueness** [1]
  the full signature determines the path (up to tree-like equivalence)

Uniqueness suggests signature inversion: recover the path from its signature.

[1] Ben Hambly and Terry Lyons. Uniqueness for the Signature of a Path of Bounded Variation and the Reduced Path Group. *Annals of Mathematics* (2010)

Truncation breaks uniqueness

- In practice, the signature is truncated as it is infinite-dimensional
- $X \mapsto S^{\leq r}(X)$ is **many-to-one**
- Existing inversion methods return a single path



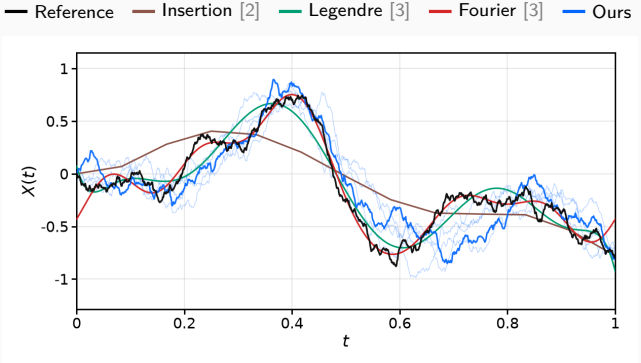
fBM ($H = 0.7$)

[2] Adeline Fermanian et al. The Insertion Method to Invert the Signature of a Path. *Recent Advances in Econometrics and Statistics*. Springer, 2024
[3] Barbora Barancikova et al. SigDiffusions: Score-Based Diffusion Models for Time Series via Log-Signature Embeddings. *ICLR*. 2025

Truncation breaks uniqueness

- In practice, the signature is truncated as it is infinite-dimensional
- $X \mapsto S^{\leq r}(X)$ is **many-to-one**
- Existing inversion methods return a single path

We should **recover a distribution of paths**, rather than a single path.



fBM ($H = 0.7$)

[2] Adeline Fermanian et al. The Insertion Method to Invert the Signature of a Path. *Recent Advances in Econometrics and Statistics*. Springer, 2024
[3] Barbora Barancikova et al. SigDiffusions: Score-Based Diffusion Models for Time Series via Log-Signature Embeddings. *ICLR*. 2025

Reformulation of signature inversion

Probabilistic signature inversion

Given $X \sim p_{\text{data}}$ and $s = S^{\leq r}(X)$, sample from the conditional distribution

$$P\big(X \mid S^{\leq r}(X) = s\big)$$

1. How do we measure the **intrinsic ambiguity** from truncation?
   $\Rightarrow$ expected distance between two samples from the conditional distribution
2. How do we **learn this distribution**?
   $\Rightarrow$ with a conditional generative model

Contribution

1. **Reformulation**: signature inversion as distribution reconstruction, not point reconstruction
2. **Theory**: a computable upper bound on the intrinsic ambiguity from truncation
3. **Experiments**: the learned sampler matches the theory, reproduces the signature, and preserves the data structure

Theory

Measuring intrinsic ambiguity

Definition 7 (Posterior-sampling reconstruction error)

For a statistic Φ, let $Q(\cdot \mid s)$ be the conditional distribution of X given $\Phi(X) = s$.

$$\mathcal{E}(\Phi) = \mathbb{E}_{X,\tilde{X}\sim Q(\cdot|\Phi(X))}\left[\int_0^T |X_t - \tilde{X}_t|^2 \, dt\right]$$

- Irreducible error of an exact posterior sampler: measuring ambiguity
- Serves as a benchmark for any learned sampler

Measuring intrinsic ambiguity

Definitions 4–6 (Truncated signature of augmented paths)

For $A \in \{\mathrm{TA}, \mathrm{TLL}\}$, $S_A^{\leq r}(X)$ keeps the terms $|I| \leq r$ of the signature of the augmented path:

$$\mathrm{TA}(X) = (t,\, X_t), \qquad \mathrm{TLL}(X) = (t,\, X_t^{\mathrm{lead}},\, X_t^{\mathrm{lag}}).$$

For $X_0 = 0$, the time coordinate makes the full signature determine the path exactly (tree-like equivalence becomes exact), and TLL also encodes the quadratic variation in its level-2 terms.

We aim to measure $\mathcal{E}(S_A^{\leq r})$, but computing it directly is challenging.

Computable upper bound via linear statistics

We consider a simpler statistic whose posterior is tractable:

Definition 9 (Linear statistic, LS)

For $r \geq 1$,

$$L^{(r)}(X) := \big(\ell_0(X), \ldots, \ell_{r-1}(X)\big)^\top, \qquad \ell_0(X) := X_T, \quad \ell_p(X) := \int_0^T t^{p-1} X_t \, dt.$$

Proposition 10 (Linear statistic upper bound)

For $r \geq 1$,

$$\mathcal{E}(S_{\mathrm{TLL}}^{\leq r}) \leq \mathcal{E}(S_{\mathrm{TA}}^{\leq r}) \leq \mathcal{E}(L^{(r)}).$$

Moreover, $\mathcal{E}(S_{\mathrm{TA}}^{\leq r}) = \mathcal{E}(L^{(r)})$ for $r \leq 2$.

Computing the bound for Gaussian processes

Contribution: computable upper bound

- For Gaussian processes, $X \mid L^{(r)}(X)$ is tractable $\Rightarrow$ $\mathcal{E}(L^{(r)})$ is computable (Prop. 11)
- Hence a computable upper bound on $\mathcal{E}(S_A^{\leq r})$ (Props. 10, 11)
- Closed form for log-GBM (Prop. 13)

Proposition 13 (Closed-form oracle error for log-GBM)

$$\mathcal{E}(L^{(r)}) = 2\, a_r\, T^2\, \mathbb{E}[\sigma^2], \qquad a_r = \frac{r}{2(2r-1)(2r+1)}.$$

Experiments

What we check

Learned sampler: flow matching (DiT backbone), conditioned on LS / TA / TLL

1. **Consistent with theory?**
   does the sampler attain the theoretical ambiguity?
2. **Faithful inversion?**
   do samples reproduce the conditioning signature?
3. **Realistic samples?**
   do samples preserve the structure of p_{data}?

Learned sampler matches the theory

LS theory $\approx$ learned LS sampler

Table 1: Posterior-sampling reconstruction error (Oracle: LS theory; LS/TA/TLL: learned sampler)

	log-GBM				drifted fBM				OU			
r	Oracle	LS	TA	TLL	Oracle	LS	TA	TLL	Oracle	LS	TA	TLL
1	1.36	1.32	1.32	1.32	1.49	1.49	1.49	1.49	1.16	1.14	1.14	1.14
2	0.54	0.55	0.55	0.55	0.66	0.66	0.66	0.67	0.50	0.51	0.51	0.51
4	0.26	0.26	0.20	0.20	0.35	0.36	0.32	0.31	0.25	0.25	0.19	0.20
6	0.17	0.17	0.14	0.14	0.25	0.27	0.23	0.25	0.17	0.17	0.14	0.15

Learned sampler matches the theory

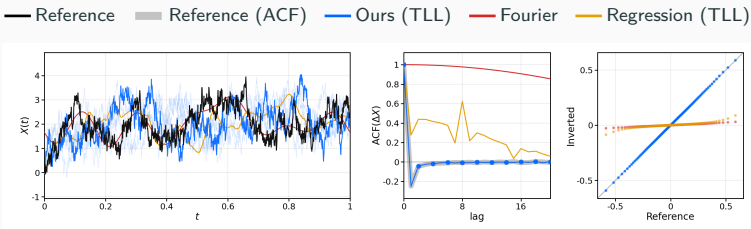
LS $\geq$ TA $\geq$ TLL, with LS $=$ TA for $r \leq 2$

Table 1: Posterior-sampling reconstruction error (Oracle: LS theory; LS/TA/TLL: learned sampler)

	log-GBM				drifted fBM				OU			
r	Oracle	LS	TA	TLL	Oracle	LS	TA	TLL	Oracle	LS	TA	TLL
1	1.36	1.32	1.32	1.32	1.49	1.49	1.49	1.49	1.16	1.14	1.14	1.14
2	0.54	0.55	0.55	0.55	0.66	0.66	0.66	0.67	0.50	0.51	0.51	0.51
4	0.26	0.26	0.20	0.20	0.35	0.36	0.32	0.31	0.25	0.25	0.19	0.20
6	0.17	0.17	0.14	0.14	0.25	0.27	0.23	0.25	0.17	0.17	0.14	0.15

Samples preserve path structure

All methods match the signature, but only ours preserves the path structure.



Samples preserve path structure

Table 2: Parameter-estimation error (ρ: Brownian correlation; bold / underline: best / second-best)

	log-GBM	drifted fBM		OU	2D log-GBM	
Method	σ err.	σ err.	H err.	σ err.	σ err.	ρ err.
Deterministic						
Fourier	1.71	1.69	0.44	1.64	1.75	0.34
Legendre	1.76	1.75	0.43	1.74	1.80	0.29
Insertion	1.93	1.93	0.46	1.90	–	–
Regression (TA)	1.73	1.72	0.33	1.72	–	–
Regression (TLL)	1.69	1.67	0.29	1.69	–	–
Probabilistic						
Ours (TA)	<u>0.22</u>	<u>0.87</u>	<u>0.05</u>	<u>0.23</u>	<u>0.24</u>	<u>0.24</u>
Ours (TLL)	**0.06**	**0.10**	**0.01**	**0.06**	**0.07**	**0.08**

Discussion

Potential use cases

- Signature-conditioned scenario generation for financial time series
- Probing what individual signature coordinates capture about a path

Limitations and future directions

- Tighter or exact bounds under truncated-signature conditioning
- Theory beyond Gaussian processes