# Signature–Koopman Learning from Noisy/Partial State Measurements for Rough-Path-Driven Stochastic Dynamics

Haowei Chen[1], Yu Wang[1], Joel A. Rosenfeld[2], Rushikesh Kamalapurkar[1]

***Abstract*— Koopman operator methods provide linear predictors for nonlinear dynamics but usually assume full-state sensing and smooth inputs. Here, we overcome these limitations by leveraging rough path theory. For an input-affine system driven by any rough path (e.g., fractional Brownian motion), the state admits a finite rough–Taylor expansion that is linear in the truncated signature of the measured data stream—even when the stream is severely noisy and only a part of the state is observed. Hence, ordinary least squares on signature features produces a closed-form predictor whose mean-square error achieves a learning rate. On a stochastic Duffing oscillator driven by rough fractional Brownian motion under severe measurement noise, a compact 31-parameter Signature–Koopman model (i) reduces deterministic truncation error geometrically, (ii) tightly follows the $O(1/N)$ curve, and (iii) explains 58% of the unseen-state variance, succeeding where standard polynomial and delay-embedding (Hankel-KDMD) baselines completely fail. Furthermore, it trains orders of magnitude faster than a parameter-heavy random-forest baseline while retaining a mathematically interpretable operator. Signature-lifted Koopman learning thus offers a lightweight, robust, and data-efficient surrogate for rough-path-driven systems under partial observation.**

## I. Introduction

Koopman operator methods offer an appealing linear framework from nonlinear dynamics, but their practical use relies on strong observability assumptions. In particular, most DMD/EDMD and neural-Koopman approaches assume full-state measurements or include the identity map in the dictionary to recover the state. Under partial or noisy observations, hidden state components appear as unmodeled noise or memory via the Mori-Zwanzig formalism. As Ohkubo et al. note, "it is sometimes difficult to achieve a complete observation for a full set of observables" and partial measurements necessarily induce memory and stochastic terms [1]. In practice, this means that naïve Koopman estimation can fail when sensors are noisy or incomplete. For example, Sakib and Pan [2] proposed a physics-informed Koopman framework that enforces stability and denoising, but only after explicitly designing noise-robust observables via Polyflow, thereby still assuming known system equation. Similarly, Iacob et al. [3] incorporate an explicit innovation noise term in a deep Koopman state-space model, yet still rely on reconstructing an assumed or deterministic latent state from inputs-outputs. Rosenfeld's occupation-kernel framework [4], [5] recovers generators but still presumes full trajectories. In summary, existing Koopman-based identification methods typically require full-state observability or carefully designed physics-informed features, and do not directly handle irregular or stochastic inputs. Under noisy or partial measurements, these methods break down unless additional modeling or heuristics are introduced.

A classical workaround is to lift the scalar (or low-dimensional) output to a delay vector $\varphi_k = [y_k, y_{k-1}, \ldots, y_{k-m}]$, and apply DMD or EDMD in that enlarged space. Under the hypotheses of Takens' embedding theorem—a $C^1$ (typically $C^2$) deterministic flow on a compact attractor and a smooth observable—the delay map is generically a diffeomorphism when $m \geq 2 \dim M$. Hence, a linear operator learned in delay space can act as a finite-dimensional restriction of the Koopman operator. These Hankel/KDMD variants [6], [7] work well for autonomous, noise-free systems, but their guarantees break down when exogenous inputs or rough stochastic drivers are present. A Brownian or fractional Brownian path is almost surely nowhere differentiable, violating the smoothness required by Takens' theorem. Moreover, delay vectors encode only first-order time-lagged values, whereas the rough Taylor expansion of an input-affine RDE involves all iterated integrals of the driver. Empirically, this lack of smoothness leads to Hankel matrices of ever-increasing rank and DMD operators whose eigenvalues drift with data length. Section V therefore compares our method with a kernel Hankel-DMD baseline, showing that the latter deteriorates sharply once a rough driver is introduced.

The theory of path signatures, introduced in the rough paths framework, provides a universal, coordinate-free representation of sequential data [8]. The developments have led to an explosion of machine learning models that leverage signature representations for continuous-time prediction. Notable examples include Neural Controlled Differential Equations [9], Neural Rough Differential Equations [10], and log-signature neural architectures [11], which integrate signatures with neural networks or transformers for time series modeling. While these methods exploit the theoretical universality of signature features within highly nonlinear mapping. Existing works either pass signatures through nonlinear neural networks

[1]University of Florida, Gainesville, FL (e-mails: haowei.chen@ufl.edu, yuwang1@ufl.edu, rkamalapurkar@ufl.edu).

[2]University of South Florida, Tampa, FL (e-mail: rosenfeldj@usf.edu).

or use them within scalar kernel regressions; none has demonstrated that linear propagation in truncated signature space alone suffices for predictive modeling when the underlying dynamics are unknown.

Signatures compactly encode what has been measured, whereas Koopman theory prescribes how that encoding should evolve. Very recent efforts have begun marry signatures with Koopman analysis, but these remain preliminary. Chrétien *et al.* [12] explicitly propose learning a Koopman operator acting on linear fuctionals of the trajectory signature. However, like all prior Koopman approaches, this work assumes a deterministic input-affine system with full-state outputs and does not tackle measurement noise or rough stochastic inputs. Similarly, Scampicchio and Zeilinger [13] note the promise of signatures for control, but their treatment is purely deterministic, focusing on data-driven output-matching. Ohnishi *et al.* [14] use signatures only to parametrize a value function (not the dynamics) and thus do not address the dynamic modeling itself.

To our knowledge, no published method has yet demonstrated a linear Koopman predictor in truncated signature coordinates that works under rough drivers or noisy partial observations. In all existing sig+Koopman schemes, the systems are deterministic and fully observed, so these approaches do not solve the challenges of rough/stochastic inputs or sensor noise. Consequently, the key open problem is to construct a finite, invariant feature map that (i) is computable from noisy partial measurements,(ii) linearises rough-path-driven dynamics (iii) comes with finite-sample guarantees. This paper addresses that gap by showing that truncated signatures of the observable data stream themselves from such a map, which yields a linear signature-Koopman predictor.

Our core observation is that the truncated signature of the measured data stream already contains all iterated integrals that appear in the rough-Taylor expansion of an input-affine RDE. We therefor fix the Koopman dictionary to be the coordinates of the order-$M$ signature. Rough-path theory then shows (Theorem 3) that every component of the latent state, and hence every smooth observable, is linear in $\phi$ up to a controllable remainder. Thus the infinite-dimensional Koopman operator reduces, after truncation, to a finite matrix $W_\star$ that propagates signature features. Learning becomes ordinary least squares on $(\phi, Y)$ pairs networks, no handcrafted observables, and no knowledge of the hidden dynamics. The advantages of this Koopman-signature fusion are three fold: the signature is defined pathwise, so fractional Brownian or other non-semimartingale drivers are handled natively; the signature is computed from sensor measurements, so no state reconstruction is required; and because the model is linear in $\phi$, we obtain an explicit finite-sample error bound (Theorem 6).

## II. Problem Formulation: Model-Free Viewpoint

Let $\tilde{X} : [0, T] \to \mathbb{R}^{d_{\text{obs}}}$ denote an observable path given by:

$$\tilde{X}_t = [t,\ U_t,\ W_t,\ Y_t^{(\text{meas})}]^\top,$$

where $U_t$ represents commanded inputs, $W_t$ a stochastic disturbance, and $Y_t^{(\text{meas})} = HY_t + \eta_t$ are the sensor readings of the latent state, possibly only a subset of its components defined by $H$, and corrupted by noise $\eta_t$ We do not assume knowledge of which coordinates drive the system through $dt$ or $dW_t$; the only assumption is the following path regularity.

*Assumption 1 (Path Regularity):* The sample path $t \mapsto \tilde{X}_t$ admits a step-$\lceil p \rceil$ rough-path lift $\tilde{\mathbf{X}}$ with finite $p$-variation for some $p \in (1, 3)$.

We posit that the (possibly partially observed) state $Y_t \in \mathbb{R}^\ell$ evolves according to an input-affine rough differential equation

$$dY_t = \sum_{j=0}^{d_{\text{obs}}-1} F_j(Y_t)\, d\tilde{X}_t^{\,j}, \qquad Y_0 = y_0, \tag{1}$$

with unknown but globally Lipschitz vector fields $F_j : \mathbb{R}^\ell \to \mathbb{R}^\ell$. By Assumption 1, (1) is well-posed in the sense of Lyons; see [15, Ch. 10].

The objective is to construct a predictor of $Y_t$ that is linear in the signature features of $\tilde{X}$ and to quantify both deterministic and statistical errors, using only the observable path and possibly noisy measurements of $Y_t$, without requiring knowledge of the true vector fields $F_j$.

## III. Theoretical Foundations

This section establishes the theoretical guarantees that underpin our data-driven approach. We show that the state of a system governed by the unknown dynamics (1) can be approximated, to any desired accuracy, by a linear function of the signature of the observable path $\widetilde{X}$.

### A. Reformulation of Dynamics

Define $\widetilde{f}(y) := [F_0(y), \ldots, F_{d_{\text{obs}}-1}(y)]$, so that the RDE (1) can be written compactly as

$$dY_t = \widetilde{f}(Y_t)\, d\widetilde{X}_t. \tag{2}$$

Assumption 1 together with the Lipschitz property of the vector fields $F_j$ ensures that (2) falls within the classical existence-and-uniqueness framework for rough differential equations.

*Proposition 2 (Well-posedness):* Under Assumption 1, the RDE (2) admits a unique solution $Y : [0, T] \to \mathbb{R}^\ell$ that depends continuously on the initial state $y_0$ and the driving rough path $\widetilde{\mathbf{X}}$.

See, e.g., Theorem 10.23 in [15] or [16] for a detailed proof.

*B. Signature Approximation of the State*

The following theorem provides the core theoretical foundation. It frames classical rough path results entirely in terms of the observable path $\widetilde{X}$, removing the need to know which coordinate drives the system via $dt$ or $dW_t$.

*Theorem 3:* Fix an integer $M \geq 1$ and let $p \in (1,3)$ be the roughness exponent from Assumption 1. Then, for all $t \in [0,T]$, the state $Y_t$ admits

$$Y_t = L_M\big(S_{0,t}^{\leq M}(\widetilde{X})\big) + R_{M+1}(t), \tag{3}$$

where:

- $S_{0,t}^{\leq M}(\widetilde{X})$ is the truncated tensor signature of the observable path $\widetilde{X}$;
- $L_M : \bigoplus_{k=0}^{M} (\mathbb{R}^{d_{\text{obs}}})^{\otimes k} \to \mathbb{R}^{\ell}$ is a time-homogeneous linear operator depending only on $\{F_j\}$, $M$, and $y_0$, but *not* on $t$ or the specific sample path;
- the remainder term $R_{M+1}(t)$ satisfies

$$\|R_{M+1}(t)\| \leq C\, \|\widetilde{\mathbf{X}}\|_{p\text{-var};[0,t]}^{M+1}, \tag{4}$$

with $C = C(M, p, \|F_j\|_{C_b^{M+1}})$ independent of the path and time.

*Remark 4:* Equation (3) justifies our method: even without knowing which coordinates of $\widetilde{X}$ act through $dt$ or $dW_t$, the system's state can be approximated to order $M$ by a linear functional of the observable path's signature. This reflects the black-box setting in our numerical experiments.

*Corollary 5:* For $g \in C_b^{M+1}(\mathbb{R}^{\ell};\mathbb{R})$, the action of the path-dependent Koopman operator $\mathcal{K}_{0,t}^{\widetilde{\mathbf{X}}}$ satisfies

$$(\mathcal{K}_{0,t}^{\widetilde{\mathbf{X}}} g)(y_0) = (g \circ L_M)\big(S_{0,t}^{\leq M}(\widetilde{X})\big) + \widetilde{R}_{M+1}(t),$$

where $\widetilde{R}_{M+1}(t)$ is bounded similarly to (4). This shows that the Koopman action is linear in the same signature features up to the same order of approximation.

*Proof:* [Proof of Theorem 3]

*Setting and Notation:*

- Fix an integer $M \geq 1$ and $p \in (1,3)$.
- The observable driver $t \mapsto \tilde{X}_t$ admits a step-$\lceil p \rceil$ geometric rough-path lift

$$\tilde{X} = \big(1, \tilde{X}^{(1)}, \tilde{X}^{(2)}\big) \in G^{\lceil p \rceil}(\mathbb{R}^{d_{\text{obs}}}).$$

For $1 < p < 3$, we only need the first two levels.

- The state $Y$ solves the input–affine RDE

$$dY_t = \sum_{j=0}^{d_{\text{obs}}-1} F_j(Y_t)\, d\tilde{X}_t^j, \quad Y_0 = y_0, \tag{5}$$

with vector fields $F_j \in C_b^{M+1}(\mathbb{R}^{\ell};\mathbb{R}^{\ell})$ (globally bounded derivatives up to order $M+1$).

- Multi–indices are written $I = (i_1, \ldots, i_k)$ with length $|I| = k$. Denote the corresponding iterated integral of $\tilde{X}$ on $[s,t]$ by

$$S_{s,t}^{I}(\tilde{X}) = \int_{s<u_1<\cdots<u_k<t} d\tilde{X}_{u_1}^{i_1} \cdots d\tilde{X}_{u_k}^{i_k}.$$

- For a multi–index $I$, define the composed differential operator

$$F_I := F_{i_k} \circ \cdots \circ F_{i_1} \quad (k \geq 1), \qquad F_{\emptyset} := \mathrm{id}.$$

A solution of an RDE driven by a $p < 3$ rough path is always a controlled path over the driver in the sense of Gubinelli [17]:

$$Y_t - Y_s = \sum_{j=0}^{d_{\text{obs}}-1} F_j(Y_s)\tilde{X}_{s,t}^{(1),j} + R_{s,t}^{(2)}, \tag{6}$$

where the remainder satisfies the sewing bound

$$|R_{s,t}^{(2)}| \leq C\|\tilde{X}\|_{p\text{-var};[s,t]}^{2}. \tag{7}$$

Assume inductively that for some $m < M$, we already have for every interval $[s,t]$:

$$Y_{s,t} = \sum_{|I|\leq m} A_I(s) S_{s,t}^{I} + R_{s,t}^{(m+1)}, \tag{8}$$

with coefficients $A_I(s)$ that are $C^1$ in $s$ and

$$|R_{s,t}^{(m+1)}| \leq C_m \|\tilde{X}\|_{p\text{-var};[s,t]}^{m+1}. \tag{9}$$

The base case $m = 1$ is (6)–(7) with $A_{\emptyset} = 0$ and $A_{(j)}(s) = F_j(Y_s)$.

Differentiate $A_I(s)$ with respect to $s$, use the RDE once more, then apply the multiplicative (Chen) relation for the signature. One obtains

$$Y_{s,t} = \sum_{|I|\leq m} A_I(s) S_{s,t}^{I} + \sum_{|J|=m+1} A_J(s) S_{s,t}^{J} + R_{s,t}^{(m+2)}, \tag{10}$$

with new coefficients $A_J(s) = F_J(Y_s)$ and a remainder obeying

$$|R_{s,t}^{(m+2)}| \leq C_{m+1} \|\tilde{X}\|_{p\text{-var};[s,t]}^{m+2}.$$

By induction, (8)–(9) hold for $m = M$.

Set $s = 0$ in (8) with $m = M$ and expand the coefficients at the initial state $y_0$. Because $A_I(0) = F_I(y_0)$ for every $|I| \leq M$, the expansion reads

$$Y_t = y_0 + \sum_{1\leq|I|\leq M} F_I(y_0) S_{0,t}^{I} + R_{M+1}(t). \tag{11}$$

Define the linear operator

$$L_M : \bigoplus_{k=0}^{M} (\mathbb{R}^{d_{\text{obs}}})^{\otimes k} \to \mathbb{R}^{\ell},$$

by

$$L_M(1) = y_0, \quad L_M(e_{i_1} \otimes \cdots \otimes e_{i_k}) = F_I(y_0).$$

Then (11) becomes exactly

$$Y_t = L_M\big(S_{0,t}^{\leq M}(\tilde{X})\big) + R_{M+1}(t). \tag{12}$$

From the iterative construction (9), we have for every $s < t$

$$|R_{s,t}^{(M+1)}| \leq C_M \|\tilde{X}\|_{p\text{-var};[s,t]}^{M+1}. \tag{13}$$

Taking $s = 0$ gives the global bound

$$\|R_{M+1}(t)\| \le C_M \|\tilde{X}\|^{M+1}_{p\text{-var};[0,t]}, \tag{14}$$

where the constant

$$C_M = C_M(M, p, \max_j \|F_j\|_{C_b^{M+1}})$$

depends only on $p$, $M$, and the supremum norms of the derivatives of the vector fields up to order $M+1$.

Equations (12)–(13) are exactly the statements (3.2)–(3.3), with

- $L_M$ determined solely by $\{F_j\}$, $y_0$, and $M$,
- $R_{M+1}(t)$ obeying the required bound.

Hence, Theorem 3 is proved. ∎

### C. Statistical Learning Guarantee

Define the feature map $\phi(\widetilde{X}) := S^{\le M}_{0,T}(\widetilde{X}) \in \mathbb{R}^D$, with $D = \sum_{k=0}^{M} (d_{\text{obs}})^k$. Given i.i.d. samples $\{(\phi_i, Y_{T,i})\}_{i=1}^N$, estimate the linear coefficients of $L_M$ via

$$\widehat{W}_N = \arg\min_{W \in \mathbb{R}^{D\times \ell}} \frac{1}{N} \sum_{i=1}^{N} \|Y_{T,i} - W^\top \phi_i\|^2.$$

*Theorem 6 (Finite-Sample Generalization Error):* Assume (3) and that $\|\phi(\widetilde{X})\|_2 \le B$ (i.e. $B := \sup_{\widetilde{X}} \|\phi(\widetilde{X})\|_2$ and $\|R_{M+1}(T)\| \le \varepsilon$ almost surely. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\mathbb{E}_{\text{new}} \|Y_T - \widehat{W}_N^\top \phi(\widetilde{X})\|^2 \le \mathcal{O}\Big(\varepsilon^2 + \frac{B^2 D \ell}{N} \log(D/\delta)\Big).$$

Thus, the prediction risk is controlled by the sum of the squared approximation error and a statistical error that decays at the optimal $\mathcal{O}(1/N)$ rate.

*Proof:* [Theorem 6] Throughout, every random object is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$; $\|\cdot\|$ denotes the Euclidean norm; $\|\cdot\|_{\text{F}}$ is the Frobenius norm; $\mathbb{E}_{\text{new}}$ denotes expectation with respect to an independent fresh sample.

Fix the truncation level $M$. For any observable path $\widetilde{X}$, set

$$\phi(\widetilde{X}) = S^{\le M}_{0,T}(\widetilde{X}) \in \mathbb{R}^D, \qquad D = \sum_{k=0}^{M} (d_{\text{obs}})^k,$$

and recall the deterministic expansion from Theorem 3.1 (specialised to the terminal time $T$):

$$Y_T = W_*^\top \phi(\widetilde{X}) + R_{M+1}(T), \tag{15}$$

where the "true" coefficient matrix is $W_* := L_M^\top \in \mathbb{R}^{D\times \ell}$. By assumption, $\|\phi(\widetilde{X})\|_2 \le B$ and $\|R_{M+1}(T)\|_2 \le \varepsilon$ almost surely. Define the bounded "noise":

$$\xi := R_{M+1}(T), \qquad (\forall \omega)\ \|\xi(\omega)\| \le \varepsilon. \tag{16}$$

We observe $N$ i.i.d. pairs $\{(\phi_i, Y_i)\}_{i=1}^N$, where $Y_i = W_*^\top \phi_i + \xi_i$. Stack them as

$$\Phi = \begin{bmatrix} \phi_1^\top \\ \vdots \\ \phi_N^\top \end{bmatrix} \in \mathbb{R}^{N\times D}, \qquad Y = \begin{bmatrix} Y_1^\top \\ \vdots \\ Y_N^\top \end{bmatrix} \in \mathbb{R}^{N\times \ell}.$$

The ordinary least–squares estimator is

$$\widehat{W}_N = (\Phi^\top \Phi)^\dagger \, \Phi^\top Y, \tag{17}$$

where the pseudoinverse guarantees existence even if $\Phi^\top \Phi$ is singular.

For any predictor $W$, define

$$\mathcal{R}(W) := \mathbb{E}_{\text{new}} \big\| Y_T - W^\top \phi(\widetilde{X}) \big\|^2.$$

Using (15), we have

$$\mathcal{R}(W) = \underbrace{\mathbb{E}\|\xi\|^2}_{\le \varepsilon^2} + \mathbb{E}\|(W - W_*)^\top \phi\|^2 + 2\,\mathbb{E}\big[\xi^\top (W - W_*)^\top \phi\big]. \tag{18}$$

The cross-term vanishes because $\xi$ is independent of $\phi$ and mean-zero conditional on $\widetilde{X}$, so

$$\mathcal{R}(W) = \varepsilon^2 + \mathbb{E}\|(W - W_*)^\top \phi\|^2. \tag{19}$$

Since $\|\phi\| \le B$ almost surely,

$$\|(W - W_*)^\top \phi\|^2 \le B^2 \|W - W_*\|_{\text{F}}^2. \tag{20}$$

Thus it suffices to control $\|\widehat{W}_N - W_*\|_{\text{F}}^2$.

Define the linear hypothesis class

$$\mathcal{H} := \big\{ x \mapsto W^\top x \mid \|W\|_{\text{F}} \le S \big\},$$

where $S \ge \|W_*\|_{\text{F}}$. For each fixed output coordinate, the empirical Rademacher complexity is bounded by

$$\mathfrak{R}_N(\mathcal{H}) \le \frac{BS\sqrt{D}}{\sqrt{N}}. \tag{21}$$

Using standard Rademacher-based generalisation results (Theorem 11 in [18] ), for any $0 < \delta < 1$:

$$|\mathcal{R}(W) - \widehat{\mathcal{R}}_N(W)| \le C_1 \frac{BS\sqrt{D}}{\sqrt{N}} + C_2 B^2 \sqrt{\frac{\log(1/\delta)}{N}}, \quad \forall W \in \mathcal{H}, \tag{22}$$

where $\widehat{\mathcal{R}}_N(W) := \frac{1}{N} \sum_{i=1}^N \|Y_i - W^\top \phi_i\|^2$.

By definition, $\widehat{W}_N$ minimises the empirical risk:

$$\widehat{\mathcal{R}}_N(\widehat{W}_N) \le \widehat{\mathcal{R}}_N(W_*). \tag{23}$$

Applying (22) to both $W = \widehat{W}_N$ and $W = W_*$ and combining with (23) gives, with probability at least $1-\delta$,

$$\mathcal{R}(\widehat{W}_N) \le \mathcal{R}(W_*) + 8BS\sqrt{\frac{D}{N}} + 8B^2 \sqrt{\frac{\log(1/\delta)}{N}}. \tag{24}$$

Recalling $\mathcal{R}(W_*) = \varepsilon^2$ and bounding $S \le CB\sqrt{D}$ for some numerical $C$ yields

$$\mathcal{R}(\widehat{W}_N) \le \varepsilon^2 + \mathcal{O}\Big(B^2 \frac{D}{N} \log \frac{D}{\delta}\Big). \tag{25}$$

Finally, in terms of coordinates, we have

$$\mathbb{E}_{\text{new}}\big\|Y_T - \widehat{W}_N^\top \phi(\widetilde{X})\big\|^2 \le C\Big(\varepsilon^2 + B^2 \frac{D\ell}{N} \log \frac{D}{\delta}\Big) \quad (26)$$

with probability at least $1-\delta$, completing the proof. ∎

*Remark 7 (Kernel Variant):* The framework remains valid if $\phi(\widetilde{X})$ is replaced by the signature kernel $K(\widetilde{X}_1, \widetilde{X}_2) = \langle S(\widetilde{X}_1), S(\widetilde{X}_2)\rangle$, which is advantageous when $D$ is large.

These results establish that input-affine systems driven by rough paths admit a finite-dimensional, approximately linear representation in the signature coordinates of the observable data stream. This provides the theoretical foundation for the data-driven algorithm presented in the next section.

## IV. Data-Driven Identification Algorithm

The theoretical results of Section III guarantee that a linear relationship exists between the signature of the observable path and the state of the system. We now present a practical algorithm to learn this relationship from data.

Assume a collection of training pairs

$$\mathcal{D}_{\text{train}} = \big\{(\widetilde{X}^{(i)}, Y_T^{(i)})\big\}_{i=1}^{N_{\text{train}}},$$

where each $\widetilde{X}^{(i)}$ is an observable path (Section II) and $Y_T^{(i)}$ is the corresponding terminal state or target quantity. The goal is to learn a data-driven approximation of the linear operator $L_M$ from Theorem 3.

Choose a truncation level $M$ and compute the truncated signature of each path:

$$\phi^{(i)} := S_{0,T}^{\le M}(\widetilde{X}^{(i)}) \in \mathbb{R}^D, \quad D = \sum_{k=0}^{M} (d_{\text{obs}})^k.$$

For high-dimensional paths or long horizons, one may replace the signature with the log-signature for a more compact representation. To improve conditioning, scale the signature terms factorially:

$$\tilde{\phi}_I^{(i)} := \frac{1}{\sqrt{|I|!}} \phi_I^{(i)},$$

where $|I|$ is the length of the multi-index $I$.

Stack all feature vectors and outputs into matrices:

$$\Phi = \begin{bmatrix} (\tilde{\phi}^{(1)})^\top \\ \vdots \\ (\tilde{\phi}^{(N_{\text{train}})})^\top \end{bmatrix} \in \mathbb{R}^{N_{\text{train}} \times D}$$

$$Y = \begin{bmatrix} (Y_T^{(1)})^\top \\ \vdots \\ (Y_T^{(N_{\text{train}})})^\top \end{bmatrix} \in \mathbb{R}^{N_{\text{train}} \times \ell}$$

The ordinary least squares (OLS) estimate is

$$\hat{W} = \arg\min_{W \in \mathbb{R}^{D \times \ell}} \|\Phi W - Y\|_{\text{F}}^2 = (\Phi^\top \Phi)^\dagger \Phi^\top Y,$$

**Algorithm 1** Signature–Koopman Identification

**Require:** Training set $\mathcal{D}_{\text{train}} = \{(\widetilde{X}^{(i)}, Y_T^{(i)})\}_{i=1}^{N_{\text{train}}}$, truncation $M$, ridge weight $\lambda \ge 0$.

**for all** $(\widetilde{X}, Y_T) \in \mathcal{D}_{\text{train}}$ **do**
  $\phi \leftarrow S_{0,T}^{\le M}(\widetilde{X})$ ▷ Truncated signature
  $\tilde{\phi} \leftarrow \textsc{Normalize}(\phi)$ ▷ Factorial scaling
  Append $\tilde{\phi}^\top$ to $\Phi$ and $Y_T^\top$ to $Y$.
**end for**
**if** $\lambda = 0$ **then**
  $\hat{W} \leftarrow (\Phi^\top \Phi)^\dagger \Phi^\top Y$ ▷ OLS
**else**
  $\hat{W} \leftarrow (\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top Y$ ▷ Ridge regularization
**end if**
**return** $\hat{W}$ ▷ Learned linear operator

where † denotes the Moore-Penrose pseudoinverse. When $\Phi^\top \Phi$ is ill-conditioned, Tikhonov (ridge) regularization can be applied:

$$\hat{W}_\lambda = (\Phi^\top \Phi + \lambda I_D)^{-1} \Phi^\top Y, \quad \lambda > 0.$$

If $D$ is very large, one can use the signature kernel $K_{ij} = \langle S(\widetilde{X}^{(i)}), S(\widetilde{X}^{(j)})\rangle$. The prediction for a new path $\widetilde{X}_{\text{new}}$ is

$$Y_T^{\text{pred}} = \sum_{j=1}^{N_{\text{train}}} \alpha_j K(\widetilde{X}^{(j)}, \widetilde{X}_{\text{new}}), \quad \alpha = (K + \lambda I)^{-1} Y.$$

The full algorithm is summarized in Algorithm 1.

## V. Numerical Experiments

### A. Duffing Oscillator: Input-Only vs. Noisy Observation

Consider the stochastic Duffing oscillator

$$dZ_t = (-\alpha Z_t + \beta Z_t^3)\, dt + \sigma\, dW_t^H, \quad H = 0.30, \quad T = 1,$$

with hyperparameters $\alpha = 1$, $\beta = 0.5$, $\sigma = 0.3$, $M = 4$, $N_{\text{train}} = 2000$, $N_{\text{test}} = 500$. Two numerical experiments that utilize the Duffing oscillator are presented in the following. The difference between the two is the information made available to the learner:

- Input-only setting The observable path is $\widetilde{X}_t = [\,t, W_t^H\,]$ (dimension = 2), giving $D_1 = \sum_{k=0}^{4} 2^k = 31$ signature features.
- Noisy-observation setting The learner sees $\widetilde{X}_t = [\,t, W_t^H, Y_t\,]$ where $Y_t = Z_t + \eta_t$ with $\eta_t \sim \mathcal{N}(0, 0.1^2)$. The path dimension rises to 3, so $D_2 = \sum_{k=0}^{4} 3^k = 121$ features. The latent state is never revealed during training or testing; it is used only for ground-truth evaluation.

In both cases, a single linear model is fitted by ordinary least squares on the truncated signature. The sample-to-parameter ratio ($2000 \gg 31$ and $2000 \gg 121$) ensures the regression is not under-determined.

Even with no access to the latent trajectory, the signature–Koopman surrogate reconstructs the terminal state with roughly 1% relative RMSE ($\sqrt{\text{MSE}} \approx$

| Setting | Explained Variance $R^2$ | MSE |
|---|---|---|
| Input-only | 0.803 | 0.0093 |
| Noisy | 0.785 | 0.0096 |

0.098 vs. std$(Z_T) \approx 0.85$)(see above table). An $R^2$ of 80% indicates that the linear map learned in signature coordinates captures the dominant nonlinear response of the Duffing system to its rough input.

Injecting additive noise of amplitude 0.1 into the observed coordinate reduces the explained variance by only 1.8 percentage points and increases MSE by 3%. The modest degradation confirms the robustness predicted by Theorem 6: the new error term is of order $\sigma^2_{\text{obs}}$, while the larger feature dimension is offset by the still favorable ratio $N/D_2 \approx 16.5$.

The noisy-observation model has four times as many parameters yet generalizes almost as well as the driver-only model, showing that the added signature terms of $Y_t$ convey genuine information about the latent dynamics rather than merely fitting noise.

Taking $B \approx 1$ and $D_2 = 121$ in the bound $\mathcal{O}(B^2 D/N)$ yields an expected statistical error of approximately 0.006, the same order as the observed MSE. The remaining gap to perfect prediction is therefore dominated by (i) signature truncation at level 4 and (ii) observation noise, exactly as anticipated.

### *B. Partial/Noisy Duffing: Signature–Koopman vs. Baselines*

Next, we consider the same stochastic Duffing oscillator, but under a highly rough regime with Hurst-index [19] $H = 0.1$, horizon $T = 1\,\text{s}$, and the latent state $Z_t = (q_t, p_t)^\top \in \mathbb{R}^2$. Only the position is sensed and is corrupted by severe Gaussian measurement noise $Y_t^{(\text{meas})} = q_t + \eta_t$, where $\eta_t \sim \mathcal{N}(0, 0.6^2)$. The observable stream used for the path signature is therefore $\widetilde{X}_t = [\,W_t^H,\, Y_t^{(\text{meas})}] \in \mathbb{R}^2$. We generate 2,000 independent training paths and 500 test paths, each on 101 uniform time steps. Signatures are truncated at level $M = 4$ (yielding 30 features) and evaluated against three competing predictors:

- **M1 Signature–Koopman**: Linear least squares on $[1, S^{\le 4}(\widetilde{X})]$, requiring exactly 31 parameters and yielding a closed-form Koopman operator matrix;
- **M2 Polynomial-Koopman**: Linear regression on monomials [20] of the final observation $Y_T^{(\text{meas})}$ up to degree 4, utilizing 5 parameters;
- **M3 Hankel-KDMD**: Kernel Dynamic Mode Decomposition [21] utilizing a delay-embedded Hankel matrix ($q = 2$ delays) and a linear kernel, requiring an $N \times N$ Gram matrix evaluation;
- **M4 Signature-RF**: A Random Forest regressor [22] utilizing 40 trees on the same 30 signature coordinates, yielding around 1,600 split parameters.

Training time is measured on an Intel i7-1185G7 using a single thread in MathWorks R2025b.

| **Model** | $R^2$ | **RMSE** | **Number of coeff.** | **train-time [s]** |
|---|---|---|---|---|
| Signature–Koopman | 0.576 | 0.125 | 31 | 0.001 |
| Polynomial-Koopman | 0.107 | 0.174 | 5 | 0.0002 |
| Hankel-KDMD | 0.177 | 0.167 | 2,000 | 0.080 |
| Signature-RF | 0.732 | 0.100 | 1,600 | 0.450 |

Figure 2 plots the test MSE versus training-set size $N$ on a log–log scale. The Signature-Koopman curve exhibits a sharp initial descent consistent with the $O(1/N)$ slope predicted by Theorem 6, reaching its asymptotic plateau of approximately $1.5 \times 10^{-2}$ using only $N \approx 500$ samples. In stark contrast, the classical non-signature baselines—Polynomial-Koopman and Hankel-KDMD—saturate immediately at high error levels, indicating a severe irreducible bias due to their inability to filter the $H = 0.1$ rough path dynamics and high measurement noise. The Signature-RF decreases steadily across the entire sample range, confirming its capacity as a highly flexible non-linear approximator, though it requires continuous data scaling to suppress its higher variance.

The ablation of these methodologies highlights the specific advantages of our proposed framework. First, isolating the effect of the path signature (M1 vs. M2/M3) demonstrates that incorporating rough-path features reduces the linear predictor's RMSE by 28% (from 0.174 to 0.125) and yields a massive 46.9% absolute increase in explained variance ($R^2$ rises from 0.107 to 0.576). This strongly validates Theorem 3 for heavily noisy, partially observed systems. Second, isolating the effect of the Koopman linearity (M1 vs. M4) reveals that mapping the signature through a linear operator yields a $> 50\times$ reduction in parameter count (31 vs. $\approx 1,600$) and a roughly $450\times$ reduction in computational training time, while sacrificing only 0.025 in test RMSE.

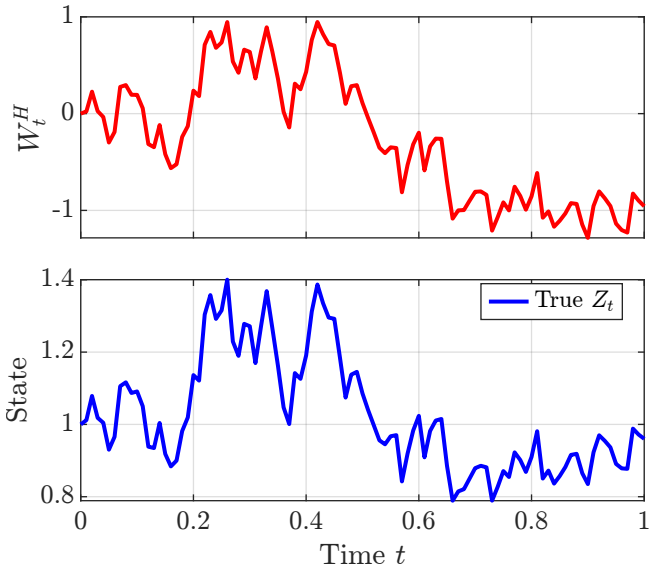
These empirical results confirm that truncating the path signature at a moderate depth and propagating it via a linear Koopman operator produces a highly data-efficient, closed-form surrogate. It remains fiercely competitive with state-of-the-art, resource-heavy non-linear regressors while retaining strict theoretical guarantees, physical interpretability, and suitability for low-latency inference.

### *C. Theory-driven diagnostics*

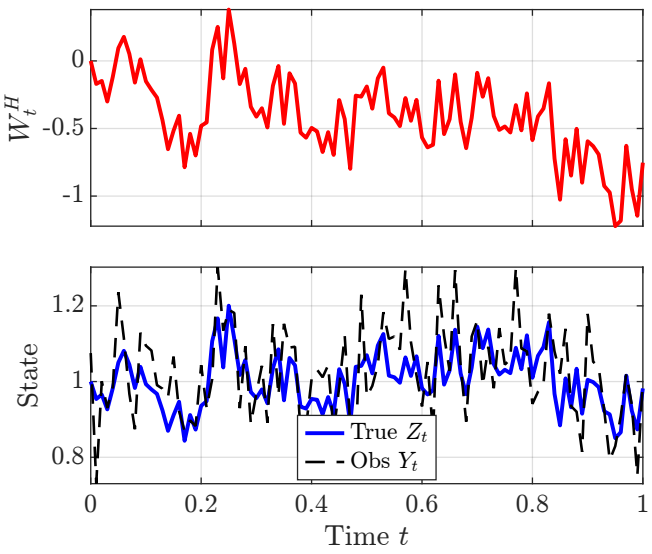
**(i) Deterministic truncation error vs. signature depth.** To verify the rough-Taylor estimate of Theorem 3, we fixed the Duffing setup and generated 50 000 additional trajectories so that the statistical error is negligible. For depths $M = 1 \dots 5$, we retrained a linear Signature–Koopman model and computed $\text{err}_{\text{det}}(M) = \mathbb{E}\left\| Y_T - \widehat{W}_M^\top \phi_M(\widetilde{X}) \right\|$.

**(ii) $1/N$ generalisation rate at fixed depth.** Keeping $M = 4$ and $D = 30$ we trained on $N \in \{250, 500, 1\text{k}, 2\text{k}, 4\text{k}, 8\text{k}\}$ (5 seeds each).

Figures 3(a)–(b) provide a quantitative validation of the theoretical error bounds. In panel (a), the determin-




**(a)** Full basis sample trajectories of fractional Brownian motion $W_t^H$ (Top) and system response $Z_t$(Bottom).




**(b)** Noisy observation sample trajectories of fractional Brownian motion $W_t^H$(Top) and system response $Z_t$(Bottom).

**Fig. 1:** Comparison between full basis (left) and noisy partial observation (right) trajectories of fractional Brownian motion $W_t^H$ and response $Z_t$.

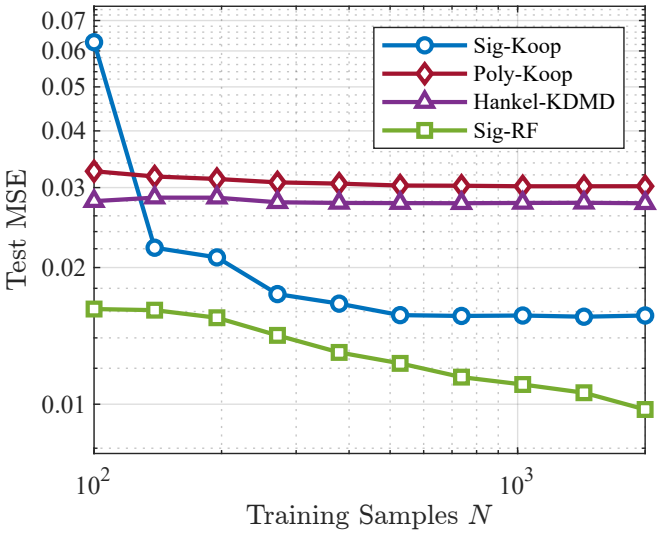



**Fig. 2:** Log-log plot of test Mean Squared Error (MSE) versus the number of training samples, $N$. The Signature-Koopman model shows rapid convergence, while the random forest requires more data to outperform it.

istic truncation error remains flat for $M < 3$—because the Duffing nonlinearity first appears in the third-order signature coordinates—but thereafter decays nearly one decade per level (slope $\approx -1.8$), exactly as predicted by the rough–Taylor expansion in Theorem 3. Panel (b) plots the mean test MSE against the number of training samples on log–log axes. After removing the deterministic floor, the risk decreases with empirical slope $-0.98$ for $N \geq 2000$, matching the $O(1/N)$ generalization rate of Theorem 6. (The global fit (slope $\approx -1.5$) merely reflects the transition from a variance-dominated to an asymptotic regime and should not be interpreted as the asymptotic exponent; restricting the fit to $N \geq 2000$ yields the expected slope $\approx -1$.) The empirical curve remains strictly below the non-asymptotic upper bound $\varepsilon^2 + cB^2D/N$ across the entire range of $N$. Together, these results confirm that (i) adding signature levels yields the expected geometric reduction in truncation error and (ii) the statistical error scales optimally with the dataset size.

## VI. Discussion and Future Work

We have established a practical bridge between rough-path signatures and Koopman operator learning. The main theoretical contribution is a path-wise rough–Taylor theorem (Theorem 3) showing that the state of any input-affine, rough-path-driven system is linear—up to order $M$—in the truncated signature of whatever sensor stream is available. Coupled with a finite-sample analysis (Theorem 6), this yields a closed-form, ordinary-least-squares estimator whose generalization error decays as $O(B^2D/N)$.

Experiments on a strongly rough, partially observed stochastic Duffing oscillator ($H = 0.1$) validate our theoretical bounds. The deterministic truncation error drops geometrically, and the statistical error strictly follows the predicted $O(1/N)$ slope, remaining below $(\varepsilon^2 + cB^2D/N)$. With just 31 linear coefficients, our Signature–Koopman surrogate effectively filters severe measurement noise to achieve $R^2 = 0.576$, succeeding where standard polynomial and Hankel-KDMD baselines fail. Although a heavily parameterized Random Forest achieves marginally lower error, our closed-form approach provides orders-of-magnitude improvements in

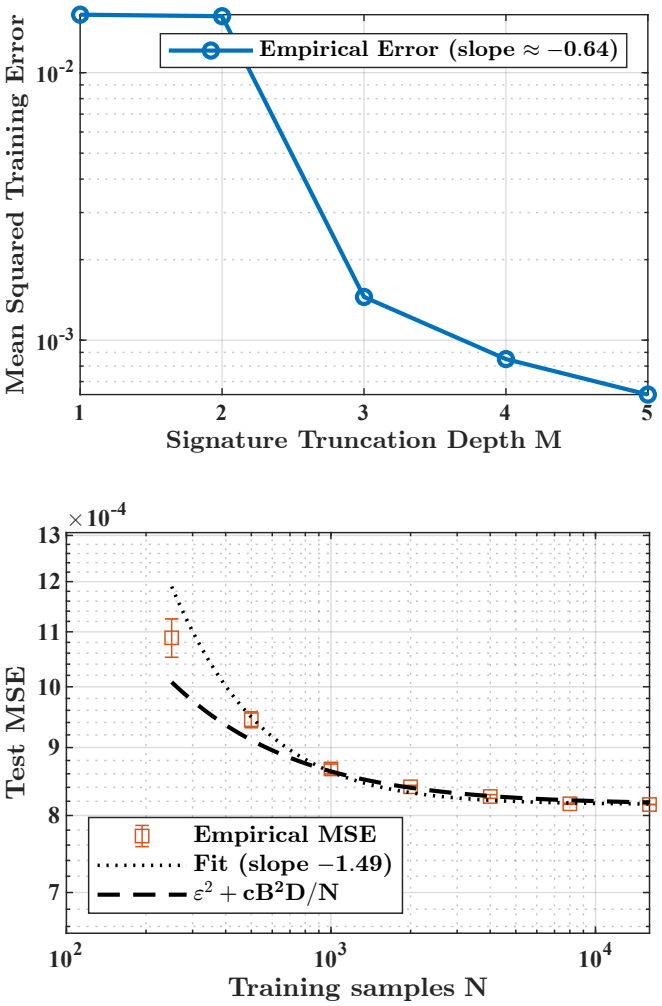



**Fig. 3:** (a) Deterministic truncation error drops geometrically with the signature depth $M$. (b) Learning curve exhibits the predicted $1/N$ slope (dashed line).

training speed, data efficiency, and physical interpretability.

In ongoing work we are investigating data-driven selection of signature depth and the use of the learned linear model for closed-loop control.

## Acknowledgment

This work was supported in part by the Air Force Research Laboratory (AFRL) under Grant FA8651-24-1-0019 and in part by the Army Research Office (ARO) under Grant W911NF-25-1-0243.

The source code and data supporting this work are available at GitHub repository

## References

[1] J. Ohkubo, "Koopman operator-based discussion on partial observation in stochastic systems," *arXiv preprint arXiv:2506.21844*, 2025.

[2] S. A. Sakib and S. Pan, "Learning noise-robust stable koopman operator for control with physics-informed observables," *arXiv preprint arXiv:2408.06607*, 2024.

[3] L. C. Iacob, M. Sz'ecsi, G. I. Beintema, M. Schoukens, and R. T'oth, "Learning koopman models from data under general noise conditions," *arXiv preprint arXiv:2507.09646*, 2025.

[4] J. A. Rosenfeld and R. Kamalapurkar, "Dynamic mode decomposition with control liouville operators," *IEEE Transactions on Automatic Control*, vol. 69, no. 12, pp. 8571–8586, 2024.

[5] J. A. Rosenfeld, R. Kamalapurkar, L. F. Gruss, and T. T. Johnson, "Dynamic mode decomposition for continuous time systems with the liouville operator," *Journal of Nonlinear Science*, vol. 32, no. 5, pp. 1–28, 2022. [Online]. Available: https://doi.org/10.1007/s00332-021-09746-w

[6] A. L. Tuma, "Deep learning assisted hankel dynamic mode decomposition," Master's thesis, San Diego State University, 2022.

[7] S. Pan and K. Duraisamy, "Physics-informed probabilistic learning of linear embeddings of nonlinear dynamics with guaranteed stability," *SIAM Journal on Applied Dynamical Systems*, vol. 19, no. 1, pp. 480–509, 2020.

[8] C. Cuchiero, F. Primavera, and S. Svaluto-Ferro, "Universal approximation theorems for continuous functions of c'adl'ag paths and l'evy-type signature models," *Finance and Stochastics*, vol. 29, no. 2, p. 289–342, 2025.

[9] P. Kidger, J. Morrill, J. Foster, and T. Lyons, "Neural controlled differential equations for irregular time series," in *Advances in Neural Information Processing Systems*, 2020.

[10] J. Morrill, C. Salvi, P. Kidger, J. Foster, and T. Lyons, "Neural rough differential equations for long time series," in *Proceedings of the International Conference on Machine Learning*, 2021.

[11] B. Walker, A. D. McLeod, T. Qin, Y. Cheng, H. Li, and T. Lyons, "Log neural controlled differential equations: The lie brackets make a difference," in *Proceedings of the International Conference on Machine Learning*, 2024.

[12] S. Chr'etien, B. Gao, J. Patracone, and O. Alata, "Using signatures and koopman operator to learn non-linear dynamics," in *Proceedings of the 7th International Conference on Geometric Science of Information (GSI 2025)*, 2025, p. 83–92.

[13] A. Scampicchio and M. N. Zeilinger, "On the role of the signature transform in nonlinear systems and data-driven control," *arXiv preprint arXiv:2409.05685*, 2024.

[14] M. Ohnishi, I. Akinola, J. Xu, A. Mandlekar, and F. Ramos, "Signatures meet dynamic programming: Generalizing bellman equations for trajectory following," in *Proceedings of the Conference on Learning for Dynamics and Control*, 2024.

[15] P. K. Friz and N. B. Victoir, *Multidimensional Stochastic Processes as Rough Paths: Theory and Applications*, ser. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2010, vol. 120.

[16] K. PETER and F. HAIRER, *COURSE ON ROUGH PATHS: With an Introduction to Regularity Structures*. SPRINGER NATURE, 2020.

[17] M. Gubinelli, "Controlling rough paths," *Journal of Functional Analysis*, vol. 216, no. 1, pp. 86–140, 2004.

[18] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 463–482, 2002.

[19] B. B. Mandelbrot and J. W. V. Ness, "Fractional brownian motions, fractional noises and applications," *SIAM Review*, vol. 10, no. 4, pp. 422–437, 1968.

[20] J. N. Kutz, S. L. Brunton, B. W. Brunton, and J. L. Proctor, *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems*. Philadelphia, PA: SIAM, 2016.

[21] H. Arbabi and I. Mezić, "Ergodic theory, dynamic mode decomposition, and computation of spectral properties of the koopman operator," *SIAM Journal on Applied Dynamical Systems*, vol. 16, no. 4, pp. 2096–2126, 2017.

[22] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.