Local Regression on Path Spaces with Signature Metrics

Christian Bayer
Weierstrass Institute

Davit Gogolashvili
Weierstrass Institute

Luca Pelizzari
University of Vienna

Abstract

We study nonparametric regression and classification for path-valued data. We introduce a functional Nadaraya-Watson estimator that combines the signature transform from rough path theory with local kernel regression. The signature transform provides a principled way to encode sequential data through iterated integrals, enabling direct comparison of paths in a natural metric space. Our approach leverages signature-induced distances within the classical kernel regression framework, achieving computational efficiency while avoiding the scalability bottlenecks of large-scale kernel matrix operations. We establish finite-sample convergence bounds demonstrating favorable statistical properties of signature-based distances compared to traditional metrics in infinite-dimensional settings. We propose robust signature variants that provide stability against outliers, enhancing practical performance. Applications to both synthetic and real-world data—including stochastic differential equation learning and time series classification—demonstrate competitive accuracy while offering significant computational advantages over existing methods.

1 INTRODUCTION

Many supervised learning problems involve *path-valued input data*—observations that take the form of sequential or temporal processes. Examples include financial asset prices evolving over time (Bouchaud et al., 2018), physiological signals such as EEG or ECG (Hannun et al., 2019; Schirrmeister et al., 2017), handwritten character trajectories (Graves et al., 2007), and human action recognition (Yang et al., 2022). Unlike the classical setting, which typically assumes fixed-dimensional vector inputs, path-valued data are theoretically infinite-dimensional, as they represent continuous functions over time intervals. In practice, such data present unique challenges arising from the observation process: they are often irregularly sampled, recorded at varying time horizons, and may exhibit strong temporal dependencies.

To address these challenges, one promising approach is the *signature transform*, first developed in Chen (1957). It provides a systematic way to extract features from sequential data by encoding a path through its iterated integrals, thereby summarizing essential information in a tensorial form. Crucially, this representation enables direct comparison of sequences of varying size and length. The signature transform further possesses several remarkable properties that make it particularly well-suited for machine learning:

- the signature naturally encodes geometrical properties of the path and is invariant under time-reparametrization;
- linear functionals of the signature are universal approximators on path space;
- the signature provides a universal representation for sequences of variable length.

Hence, the use of signatures in statistical learning has received considerable attention. While we refer to the review article by Lyons and McLeod (2022) for a broad overview, we focus on discussing the most directly related contributions. Within the machine learning literature, signatures are most commonly regarded as a feature extraction technique, a perspective that has been applied in a wide range of practical applications. Extending this view, Morrill et al. (2020) introduce the Generalised Signature Method, which provides a unifying framework for variations of the signature transform in multivariate time series analysis.

The literature most closely related to our work concerns signature methods for functional regression. In the parametric setting, linear functional regression with signatures—both with and without regularization—has been well-established in the literature (Fer-

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

manian, 2021; Gu et al., 2024; Fermanian, 2022; Bleistein et al., 2023; Cohen et al., 2023; Guo et al., 2025; Bayer et al., 2025a). There, the regression functional is assumed to be linear in the signature with finitely many unknown coefficients, a representation motivated by the Stone–Weierstrass theorem. Fully nonparametric approaches are represented by works on signature kernels (Király and Oberhauser, 2019; Chevyrev and Oberhauser, 2022; Lee and Oberhauser, 2025; Schell and Alaifari, 2023; Horvath et al., 2023; Bayer et al., 2025b). While signature kernels can be computed efficiently (Király and Oberhauser, 2019; Salvi et al., 2021; Lemercier et al., 2024; Tóth et al., 2025), these methods inherit the well-known scalability limitations of kernel learning, in particular the computational burden associated with inverting large Gram matrices. The same limitation applies to related Bayesian approaches, such as Gaussian process regression with signature kernel covariance functions (Toth and Oberhauser, 2020). To address the scalability issue, Lemercier et al. (2021a) proposed a sparse variational inference framework for Gaussian processes. Alternatively, signatures can also be used as a feature within a deep learning pipeline (Kidger et al., 2019; Walker et al., 2024; Moreno-Pino et al., 2024).

The field of nonparametric statistics for functional data represents a well-established area of research. The work of Ferraty and Vieu (2006) provides a comprehensive treatment of kernel-based methods for functional data analysis, establishing the theoretical foundations for nonparametric regression when the input space is infinite-dimensional. Subsequent developments include convergence rate analysis (Lian, 2012; Meister, 2016), variable selection methods (Aneiros et al., 2022; Shang, 2014), dependent data analysis (Masry, 2005) and extensions to more complex functional structures (Selk and Gertheiss, 2023).

Within the framework of kernel-based functional regression, the choice of *semi-metric* plays a crucial role in determining both theoretical properties and practical performance. Classical approaches typically rely on L^p metrics or Hölder norms. However, these standard choices often fail to reflect the intrinsic geometric structure of path-valued data. In fact, the associated topologies invariably produce *small-ball probabilities* of exponential form—whether based on the supremum norm, L_p, or Hölder norms (Ferraty and Vieu, 2006)—which limits their ability to enhance concentration properties. This motivates the exploration of *semi-metrics*, which can be designed specifically for sequential data and provide a more natural framework for comparing paths.

Contributions. In this paper, we propose a new estimator that leverages signature transforms within the classical local kernel (or Nadaraya–Watson) regression framework. Our contributions are twofold: (i) we establish rigorous finite-sample convergence guarantees for the proposed estimator, addressing a gap in the theoretical understanding of nonparametric methods based on signatures, and (ii) we provide an efficient and straightforward implementation that avoids the computational and scalability challenges commonly associated with signature kernel approaches, demonstrating its practical applicability on real-world datasets.

The remainder of the paper is organized as follows. Section 2 presents the background material along with the proposed estimator. Section 3 is devoted to our main theoretical results, including the convergence analysis for signature-based Nadaraya-Watson estimators and the derivation of convergence rates. Section 4 provides detailed experimental validation on synthetic and real-world datasets.

2 NOTATIONS AND BACKGROUND

We consider the problems of regression and classification for data $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is a (infinite-dimensional) path-space, and $\mathcal{Y}$ is some finite-dimensional Euclidean space. We assume that the data samples are drawn from a joint probability measure P on $\mathcal{X} \times \mathcal{Y}$.

Our central object of interest is the *regression function*, given by

$$F(x) = \mathbb{E}[Y|X = x] = \int y dP(y|x), \quad x \in \mathcal{X}, \tag{1}$$

where $P(\cdot|x)$ denotes the conditional distribution of Y given $X = x$. For instance, in financial modeling one may view the data $X = (X_t)_{t\geq 0} \in \mathcal{X}$ as the dynamics of an asset price. Then, for any *option payoff* $Y = \Psi(X)$, the conditional expectation (1) represents its fair price.

For the classification problem, we consider *categorical* responses for the path-valued data, that is we are interested in conditional probabilities

$$p_g(x) = \mathbb{P}[Y = g|X = x], \quad (x, g) \in \mathcal{X} \times \mathcal{Y}. \tag{2}$$

A typical classification example is the handwriting recognition problem, where $\mathcal{Y}$ is the alphabet $\{A, B, \ldots, Z\}$, and the data $X \in \mathcal{X}$ may be seen as handwritings of such letters - represented as paths in $\mathbb{R}^2$. The function p_g then assigns the probability of such a handwriting to correspond to the letter $g \in \mathcal{Y}$.

Both the regression and classification problems reduce to learning a conditional expectation, so their treatment follows the same principle, which we now briefly outline. Assume we have i.i.d. (independent and identically distributed) data of input-output pairs

$$(X^{(1)}, Y^{(1)}), \dots, (X^{(M)}, Y^{(M)}) \overset{\text{i.i.d.}}{\sim} P.$$

Motivated by the classical Nadaraya-Watson estimate (Nadaraya, 1964; Watson, 1964), and in particular the functional extensions thereof (Ferraty and Vieu, 2006), we consider the estimator for the regression (1)

$$\widehat{F}(x) = \frac{\sum_{i=1}^{M} Y^{(i)} K(h^{-1}\varrho(x, X^{(i)}))}{\sum_{j=1}^{M} K(h^{-1}\varrho(x, X^{(j)})}, \tag{3}$$

where ϱ is a semi-metric[1] on the path-space $\mathcal{X}$, $K : \mathbb{R} \to \mathbb{R}_+$ some asymmetric kernel, and $h = h(M)$ is strictly positive. The same estimator can be applied to the classification problem (2) by replacing Y with $1_{\{Y=g\}}$, since the latter can be viewed as a regression problem with target $p_g(x) = \mathbb{E}[1_{\{Y=g\}}|X = x]$. It is well-known in the literature (see, e.g. (Ferraty and Vieu, 2006, Chapter 13)), that compared to the finite-dimensional case, the choice of the semi-metric is very sensible from both a theoretical and practical point of view. In the following section, we introduce variants of the estimator (3), where the semi-metric ϱ is defined via the *signature transform* of the data, $\mathcal{X} \ni x \mapsto \mathrm{Sig}(x)$; see (8)–(9).

3 LOCAL REGRESSION WITH SIGNATURES

This section presents our main theoretical contributions. We establish finite-sample convergence bounds for the estimator (3) under general semi-metrics, then specialize to signature-based distances that exploit the geometric structure of rough paths. Readers unfamiliar with signatures may consult Appendix B.1 for the necessary background from rough path theory.

3.1 Convergence analysis

As is customary in nonparametric regression, we impose smoothness constraints on the regression function F. The choice of these constraints is particularly delicate in the infinite-dimensional setting, where the topology of the underlying space $\mathcal{X}$ plays a crucial role in determining both the convergence rates and the practical performance of the estimator.

[1]Satisfying all properties of a metric except for point-separation, i.e. $\varrho(x, y) = 0 \not\Rightarrow x = y$.

Definition 3.1 *Let $\beta \in (0, 1]$ and L be any positive constant. For any semi-metric ϱ on $\mathcal{X}$, we denote by $\mathcal{F}_\beta^\varrho$ the Hölder class of functionals $F : \mathcal{X} \to \mathbb{R}$ that satisfy the condition*

$$|F(x) - F(x')| \le L\varrho(x, x')^\beta, \tag{4}$$

for all $x, x' \in \mathcal{X}$.

Remark 3.2 *It is important to note that our convergence analysis applies to smoothness levels $\beta \le 1$. Since the domain $\mathcal{X}$ is not a vector space in general, extending to higher smoothness levels presents significant challenges due to the absence of a natural notion of differentiation.*

A fundamental quantity in the convergence analysis is the *concentration function*, which measures how the data concentrates around a given point in the metric space.

Definition 3.3 *For a random variable X taking values in a semi-metric space $\mathcal{X}$, with semi metric ϱ, we define the concentration function*

$$\phi_x^\varrho(h) = \mathbb{P}[\varrho(X, x) \le h], \quad h \ge 0, \tag{5}$$

where the probability is with respect to the distribution of X and $x \in \mathcal{X}$.

The behaviour of $\phi_x^\varrho(h)$ as $h \downarrow 0$ determines the convergence rates of our estimator. Intuitively, slower decay of $\phi_x^\varrho(h)$ indicates that the data is less dispersed around x, leading to better statistical performance. This behaviour is intimately connected to the geometry of the path space and the choice of distance function.

In the classical finite-dimensional setting, when $\mathcal{X} = \mathbb{R}^d$, the concentration function typically exhibits polynomial decay $\mathcal{O}(h^d)$ regardless of the choice of the metric (as in finite-dimensional spaces all norms are equivalent). However, in infinite-dimensional spaces, the situation is more delicate and depends heavily on the choice of metric.

The choice of the semi-metric ϱ in Definition 3.1 is of paramount importance, as it directly influences both the class of admissible functions $\mathcal{F}_\beta^\varrho$ and the small-ball probability behaviour ϕ^ϱ that governs the convergence rates. For brevity, we will often write $\mathcal{F}_\beta = \mathcal{F}_\beta^\varrho$ and $\phi_x = \phi_x^\varrho$ whenever the choice of ϱ is fixed and clear from the context.

In the setting of path-valued data, we propose using signature-based distances in Section 3.2, which naturally respect the geometric structure of the underlying paths. Before doing so, we establish the fundamental

convergence rates of the Nadaraya–Watson estimator with respect to *any* semi-metric ϱ on $\mathcal{X}$.

Theorem 3.4 *Let $Y \in [-R, R]$ and let $F \in \mathcal{F}_\beta$ with smoothness parameter $\beta \in (0, 1]$. Consider the estimator $\widehat{F}$ defined in (3), and assume that the kernel K is compactly supported and satisfies*

$$b1_{[0,1]} \leq K \leq B1_{[0,1]}, \tag{6}$$

for some constants $0 < b \leq B < \infty$.

Let $\delta \in (0, 1)$ and suppose that M satisfies

$$M \geq \frac{16B \log(6/\delta)}{b\phi_x(h)}.$$

Then, with probability at least $1-\delta$ the following bound holds

$$|\widehat{F}(x) - F(x)| \leq \frac{B}{b}h^\beta + 8R\sqrt{\frac{2B\log(6/\delta)}{b\phi_x(h)M}}. \tag{7}$$

The proof of the theorem can be found in Appendix A. Theorem 3.4 provides a finite-sample bound of the pointwise error for the Nadaraya-Watson estimator, which decomposes the estimation error into two fundamental components: the *bias* term that is directly controlled by the Hölder parameter β from Definition 3.1, reflecting how the local smoothness of F around the target point x affects the estimation quality. And the *variance* term $\mathcal{O}\left((\phi_x(h)M)^{-1/2}\right)$, that captures the stochastic fluctuations due to finite sample size M.

In the Euclidean case, when $\mathcal{X} \subseteq \mathbb{R}^d$, the concentration function exhibits polynomial decay $\phi_x(h) \sim Ch^d$, for some constant $C > 0$. Assuming an optimal choice of bandwidth $h \sim M^{-1/(2\beta+d)}$ that balances the bias-variance trade-off, Theorem 3.4 yields the classical nonparametric rates of convergence $M^{-\beta/(2\beta+d)}$, that is known to be optimal in the minimax sense (Stone, 1980).

However, in infinite-dimensional spaces, the situation becomes significantly more delicate and the convergence behavior fundamentally changes. The concentration function typically exhibits exponential behavior $\phi_x(h) \sim C\exp\left(-ch^{-\gamma}\right)$.

Gaussian processes provide an illuminating example. For instance, for fractional Brownian motion $X^H = (X_t^H : 0 \leq t \leq 1)$ we have $\phi_x(h) \sim C\exp(-h^{-2/(2H-\beta)})$ when ϱ is chosen to be the β-Hölder distance (Li and Shao, 2001).

Exponential decay of the concentration function leads to convergence rates that are fundamentally slower than their finite-dimensional counterparts. Specifically, under regularity conditions on the regression function $F \in \mathcal{F}_\beta$ and appropriate choice of bandwidth, one can achieve slow logarithmic type rates $\mathcal{O}(\log(M)^{-2\beta/\gamma})$, known to be optimal in the minimax sense (Meister, 2016).

In the following subsection, we analyze the concentration function behavior for two specific choices of signature-based distances that are particularly relevant for rough path data.

3.2 Signature semi-metrics on path-spaces

Our main focus in the applications below is on *path-valued data*, that is, on spaces $\mathcal{X}$ consisting of paths $x : [0, T] \to E$, where the state space is typically the Euclidean space $E = \mathbb{R}^d$. In this section, we propose a canonical metric choice induced by the signature transform, which offers several theoretical and practical advantages over conventional distances on path spaces.

For simplicity of exposition, we assume in this section that $\mathcal{X} = C^1([0, T], \mathbb{R}^d)$, whereas a more general construction for *rougher data*—namely $\mathcal{X} = C^{p-var}([0, T], \mathbb{R}^d)$ with $p \geq 1$—is discussed in Appendix B.1. The signature of a path $x \in \mathcal{X}$ is given as a sequence of tensors

$$\mathrm{Sig}(x) = \left(1, \mathrm{Sig}(x)^{(1)}, \ldots, \mathrm{Sig}(x)^{(k)}, \ldots\right) \in \prod_{k\geq 0}(\mathbb{R}^d)^{\otimes k},$$

consisting of iterated integrals

$$\mathrm{Sig}(x)^{(k)} = \left(\int_0^T \int_0^{t_k} \cdots \int_0^{t_2} dx_{t_1}^{i_1} \cdots dx_{t_k}^{i_k}\right)_{i_1,\ldots,i_k \in [d]},$$

where $[d] = \{1, \ldots, d\}$, see also Definition B.2.

Among the many fascinating algebraic and analytical properties of the signature transform, some of which we summarize in Appendix B.1, the one most relevant for this article is that the sequence $\mathrm{Sig}(x)$ can be regarded as an *encoding* or *description* of the trajectory. Indeed, the transform $x \mapsto \mathrm{Sig}(x)$ is injective up to some equivalence class $\sim$ (see Appendix B.1 for more details), so that the sequence $\mathrm{Sig}(x)$ uniquely characterizes the underlying path x. Such equivalence classes become trivial once paths are augmented with time, $x_t \mapsto (t, x_t)$; see (24) and the preceding discussion. At the cost of increasing the dimension by one, we henceforth assume that

$$\mathcal{X} = \widehat{C}^1 = \{x_t = (t, x_t) : x \in C^1([0, T], \mathbb{R}^d)\}.$$

Injectivity of the signature transform allows us to introduce a *Euclidean-like* distance between paths

$$\varrho^{\mathrm{Sig}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+, \quad \varrho^{\mathrm{Sig}}(x, y) = \|\mathrm{Sig}(x) - \mathrm{Sig}(y)\|, \tag{8}$$

where

$$\|\mathbf{a}\| = \sqrt{\sum_{k\geq 0} \|\mathbf{a}^{(k)}\|^2_{(\mathbb{R}^d)^{\otimes k}}}, \quad \mathbf{a} \in \prod_{k\geq 0} (\mathbb{R}^d)^{\otimes k},$$

see Lemma B.6. In words, the distance between two data points x and y is measured by comparing their signatures in the *extended tensor algebra* $\mathcal{T} = \prod_{k\geq 0}(\mathbb{R}^d)^{\otimes k}$. In Appendix B.1 we provide a more detailed introduction to the algebraic structure of $\mathcal{T}$, including its product $\otimes$ and addition $+$, as well as further details on its Hilbert space structure.

Remark 3.5 *A more conventional distance on C^1 is given by the 1-variation (or length) of the difference of two paths, namely*

$$\varrho^1(x,y) = \|x-y\|_{1\text{-}var}, \quad \|x\|_{1\text{-}var} = \int_0^T |\dot{x}_t|\, dt.$$

We know $\|\mathrm{Sig}(x)^{(k)}\|_{(\mathbb{R}^d)^{\otimes k}} \leq \frac{\|x\|^k_{1\text{-}var}}{k!}$ from Lemma B.4, so that the distance in (8) is finite.

A key feature of (8) on the infinite-dimensional space $\mathcal{X}$ is that it naturally admits finite-dimensional projections, obtained by truncating the sequence $\mathrm{Sig}(x)$ at some tensor level N. Supported by the factorial decay noted in Remark 3.5, for N large enough the distance (8) is well-approximated by its truncated version

$$\varrho^{\mathrm{Sig}}_{\leq N}(x,y) = \sqrt{\sum_{n=0}^{N} \left\|\mathrm{Sig}(x)^{(n)} - \mathrm{Sig}^{(n)}(y)\right\|^2_{(\mathbb{R}^d)^{\otimes n}}}. \tag{9}$$

Remark 3.6 *Both the truncated and untruncated signature distances are well-supported by publicly available open-source libraries, such as iisignature (Reizenstein and Graham, 2020) or roughpy (Morley and Lyons, 2024). In particular, for the untruncated distance one can exploit its relation to the* signature kernel

$$\varrho^{\mathrm{Sig}}(x,y)^2 = k(x,x) - 2k(x,y) + k(y,y),$$

where k is obtained by solving a Goursat-type PDE; (Salvi et al., 2021).

Returning to the local regression analysis from Section 3.1, we now derive convergence rates for the truncated signature distance. Before doing so, we show in the following lemma that in both the truncated and untruncated cases, the rate of convergence is always at least as fast as under the 1-variation metric; the proof can be found in Appendix B.2.

Lemma 3.7 *For any $\mathcal{R} > 0$ and random variable X in $\mathcal{X}_\mathcal{R} = \{x \in \mathcal{X} : \|x\|_{1-var} \leq \mathcal{R}\}$, we have*

$$\phi_x^{\varrho^{\mathrm{Sig}}_{\leq N}}(h) \geq \phi_x^{\varrho^{\mathrm{Sig}}}(h) \geq \phi_x^{\varrho^1}(Ch), \quad \forall x \in B_\mathcal{R},$$

for some constant $C > 0$.

To derive convergence rates, we recall from Appendix B.1 that the signature takes values in a free nilpotent Lie group $\mathcal{G} \subset \mathcal{T}$, which will be crucial for the following assumption.

Assumption 3.8 *We assume that X is a random variable taking values in $\mathcal{X}$, and its truncated signature*

$$\mathrm{Sig}(X)^{\leq N} = \left(1, \mathrm{Sig}^{(1)}(X), \dots, \mathrm{Sig}^{(N)}(X)\right)$$

admits a density function p with respect to the Haar measure of the Lie group $\mathcal{G}^{\leq N}$, see Definition B.7, which is bounded away from zero, i.e., $p(\mathbf{g}) \geq c > 0$ for all $\mathbf{g} \in \mathcal{G}^{\leq N}$.

Example 1 While we restrict to smooth random paths X here for simplicity, Assumption 3.8 remains reasonable in rougher frameworks. For instance, it has been shown to hold for Brownian motion $X = B$ already in Kusuoka and Stroock (1987), which is relevant for our application in Section 4.1, and has further been extended to fractional Brownian motion $X = B^H$ with $H > 1/4$ recently in Baudoin et al. (2020), relevant for the application in Appendix C.2.

Proposition 3.9 *Under Assumption 3.8, the small-ball probability with respect to the truncated signature distance (9) has at most polynomial decay, in the sense that for any $N \in \mathbb{N}$*

$$\phi_x^{\varrho^{\mathrm{Sig}}_{\leq N}}(h) = \mathbb{P}\left[\varrho^{\mathrm{Sig}}_{\leq N}(X,x) \leq h\right] \geq Ch^{\nu(N)}, \quad x \in \mathcal{X},$$

where $C > 0$ is constant and

$$\nu(N) = \sum_{n=1}^{N} \frac{1}{n} \sum_{\ell | n} \mu\left(\tfrac{n}{\ell}\right) d^\ell,$$

and $\mu(\cdot)$ denotes the Möbius function, and the inner sum is taken over divisors of n.

Remark 3.10 *While the rigorous proof is deferred to Appendix B.2, we briefly indicate why such a result is natural. Owing to the density assumption, the small-ball probability is bounded below by the volume of balls B_h in $\mathcal{G}^{\leq K}$. We will see that these volumes scale as $h^{\dim(\mathfrak{g}^{\leq K})}$, where $\mathfrak{g}^{\leq N}$ is the Lie algebra associated with $\mathcal{G}^{\leq N}$, whose homogeneous dimension is in fact given by $\dim(\mathfrak{g}^{\leq N}) = \nu(N)$, see (Reutenauer, 2003, Theorem 6).*

As a consequence, we obtain the following nonparametric convergence rate for our estimator (3) with respect to the truncated signature distance, the proof can be found in Appendix B.2.

Corollary 3.11 *Let Y be a fixed random variable in $[-R, R]$ and assume*

$$F(x) = \mathbb{E}[Y|X = x] \in \mathcal{F}_\beta^\varrho, \quad \varrho = \varrho_{\leq N}^{\mathrm{Sig}},$$

for some $N \in \mathbb{N}$. Suppose that the kernel K satisfies (6) and Assumption 3.8 holds true, and set

$$h = \left(\frac{\log(6/\delta)}{M}\right)^{1/(2\beta+\nu(N))} \tag{10}$$

where $\delta \in (0,1)$. Then, for M large enough, with probability at least $1-\delta$

$$|\widehat{F}(x) - F(x)| \leq C\left(\frac{\log(6/\delta)}{M}\right)^{\frac{\beta}{2\beta+\nu(N)}},$$

where C is independent of M and δ.

For fixed N, the previous result yields Euclidean-type nonparametric rates in the infinite-dimensional space $\mathcal{X}$, governed by the effective dimension $\nu(N)$. We emphasize once more that the choice of semi-metric not only affects the convergence rate, but also determines the class of admissible functions $\mathcal{F}_\beta^\varrho$, which is expected to be much smaller compared to the full signature distance. While it might not always be justified that the underlying functions lie in $\mathcal{F}_\beta$, we observe excellent performance when using the truncated signature distance in all our applications.

Remark 3.12 *While we have only considered smooth data in this section, i.e. $\mathcal{X} = \widehat{C}^1$, typical stochastic processes such as Brownian motion have more irregular sample paths. A more suitable framework is the space of paths of finite p-variation with $p \geq 2$; see Definition B.1. As outlined in Remark B.3, by exploiting rough path theory (Lyons, 1998), the signature remains well defined using Stratonovich iterated integrals (Friz and Victoir, 2010, Chapter 13)*

$$\mathrm{Sig}(\widehat{B})^{i_1\cdots i_n} = \int_0^T \int_0^{t_n} \cdots \int_0^{t_2} \circ d\widehat{B}_{t_1}^{i_1} \cdots \circ d\widehat{B}_{t_n}^{i_n}.$$

The induced distance ϱ_{Sig}—and the theory developed in this article—remains valid on rough path spaces.

3.3 Robustification

In our numerical experiments in Section 4, we observe that the estimators $\widehat{F}$ are not robust to samples yielding unusually large signature values, leading to outliers in the predictions and unstable performance.

For Brownian signatures, see also Remark 3.12, such outliers can occur for samples that produce large signature values. In this case, the signature distance to a "typical" signature sample becomes very large, and consequently the estimator (3) predicts a value close to zero. This phenomenon is not difficult to anticipate, say for $T = 1$, since the signature contains the powers

$$\int_0^1 \int_0^{t_n} \cdots \int_0^{t_2} \circ d\widehat{B}_{t_1}^2 \cdots \circ d\widehat{B}_{t_n}^2 = \frac{B_1^n}{n},$$

where $B_1 \sim \mathcal{N}(0,1)$. With small probability (Gaussian tail-estimates), these entries can become arbitrarily large whenever $B_1 \gg 1$. We illustrate and further comment on this observation in Figure 1 in Section 4.1.

We found that this issue can be addressed by adopting the robust signature proposed in Chevyrev and Oberhauser (2022), which we summarize below; further details are provided at the end of Appendix B.2, an explicit construction of Λ is provided in Example 2.

The robust signature, following Chevyrev and Oberhauser (2022), is defined by $\mathrm{RSig}(x) = \Lambda \circ \mathrm{Sig}(x)$, where Λ is a tensor normalization

$$\Lambda : \mathcal{T} \longrightarrow \{\mathbf{a} \in \mathcal{T} : \|\mathbf{a}\| \leq R\},$$

for some fixed $R > 0$. The map Λ is required to be continuous and injective in order to preserve the structural properties of the signature transform.

4 APPLICATION

In this section, we test our estimator (3) in two applications. First, we learn the solution map of stochastic differential equations as a functional of the driving noise, formulated as a nonparametric regression problem. Second, we apply the method to classification tasks on various real-world datasets consisting of sequential data, which we interpret as piecewise linear paths.

All signature computations are performed using the `iisignature` library (Reizenstein and Graham, 2020), which provides efficient implementations of signature algorithms with computational complexity $\mathcal{O}(Ld^N)$ where L is the path length, d is the dimension, and N is the truncation level. The code to reproduce all experiments is publicly available at `https://github.com/Davit1123/nw-signature`.

Our basic method, which we call `Sig`, uses the estimator defined in equation (3) with the truncated signature-based[2] semi-metric from equation (9).

[2]We also tested replacing the standard signature with

M_{tr}	RSig$^{\leq 2}$	Sig$^{\leq 2}$	RSig$^{\leq 3}$	Sig$^{\leq 3}$	RSig$^{\leq 4}$	Sig$^{\leq 4}$	2-var	4-var	L^1	L^2	Sup
8	0.115	0.146	0.114	0.163	0.114	0.172	0.206	0.323	0.197	0.190	0.177
16	0.088	0.130	0.090	0.150	0.090	0.160	0.190	0.161	0.211	0.219	0.166
32	0.060	0.111	0.061	0.128	0.061	0.138	0.179	0.213	0.183	0.164	0.154
64	0.062	0.126	0.126	0.127	0.071	0.135	0.180	0.142	0.175	0.160	0.139
128	0.054	0.104	0.051	0.111	0.051	0.119	0.164	0.129	0.141	0.141	0.121
256	0.047	0.056	0.042	0.060	0.039	0.069	0.167	0.128	0.124	0.121	0.104
512	0.042	0.055	0.033	0.064	0.032	0.061	0.163	0.124	0.111	0.106	0.093
1024	0.041	0.051	0.029	0.050	0.027	0.050	0.164	0.093	0.104	0.097	0.088
2048	0.040	0.050	0.026	0.048	0.025	0.048	0.165	0.087	0.100	0.091	0.085
4096	0.039	0.040	0.023	0.035	0.021	0.042	0.164	0.080	0.098	0.087	0.083
8192	0.038	0.040	0.021	0.034	0.019	0.041	0.165	0.074	0.098	0.084	0.083
time	15.29	0.86	18.68	1.25	23.94	2.20	1840.87	1704.02	24.85	22.52	33.54

Table 1: SDE regression accuracy (RMSE) on the testing data. The last row corresponds to the evaluation time (in seconds) required for the whole procedure for the largest training sample size $M_{tr} = 8192$. Note that we do not include 1-var here, since almost surely $\|B\|_{1-var} = +\infty$.

Throughout the experiments involving local regression, we use the standard Gaussian kernel.

The RSig method uses robust signature features to improve stability and discriminative power. For each time series, we compute truncated signatures up to level N and apply a robust transformation $\Lambda \circ \mathrm{Sig}$, where the normalization map Λ rescales each signature tensor according to its magnitude (see Appendix B.1, Example 2). This rescaling reduces sensitivity to outliers by dampening the influence of large signature components (Chevyrev and Oberhauser, 2022).

We select hyperparameters (bandwidth h, robust parameters C and a) via cross-validation. The signature level is set according to the time series dimension, and the bandwidth and robustness parameters are selected from predefined grids.

4.1 Learning the solution map of SDEs

Let $(B_t)_{t\in[0,T]}$ be an m-dimensional Brownian motion, and consider its time-augmentation $\widehat{B}_t = (t, B_t)$. We are interested in the *Itô-map* $\widehat{B} \mapsto Z_T$, where

$$Z_0 = z_0, \quad dZ_t = b(Z_t)dt + \sigma(Z_t)dB_t, \quad 0 < t \leq T,$$

with coefficients $b : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : [0,T] \times \mathbb{R}^d \to \mathbb{R}^{d\times m}$ sufficiently regular such that a unique strong solution exists, see, e.g., (Protter, 2005, Chapter 3, Theorem 7).

For the numerical experiments in this section, we fix $m = d = 1$ and consider the smooth coefficients

$$b(x) = -x^p, \quad \sigma(x) = x\cos(x), \quad p \in \mathbb{N},$$

and choose $p = 5$. For some $M \in \mathbb{N}$, we draw $B^{(1)}, \dots, B^{(M)}$ independent Brownian sample paths on some grid $\{t_0, \dots, t_L\}$ in $[0,T]$, and denote by $Y^{(m)}$ the terminal values $Y^{(m)} = Z_T^{(m)}$, obtained using an Euler-Mayurama scheme. We split the data into disjoint training and testing sets $I_{tr} \dot\cup I_{te} = \{1, \dots, M\}$ and for $m \in I_{te}$ consider

$$\widehat{Y}^{(m)} = \frac{\sum_{i\in I_{tr}} Y^{(i)} K\Big(h^{-1}\varrho(\widehat{B}^{(m)}, \widehat{B}^{(i)})\Big)}{\sum_{j\in I_{tr}} K\Big(h^{-1}\varrho(\widehat{B}^{(m)}, \widehat{B}^{(j)})\Big)}. \tag{11}$$

In Figure 1 we plot the testing data $\{(Y^{(m)}, \widehat{Y}^{(m)}) : m \in I_{te}\}$, choosing $M = 2^{13}$ and a $90\% - 10\%$ split, i.e. $M_{tr} = |I_{tr}| = 0.9 \times M$ and $M_{te} = |I_{te}| = 0.1 \times M$. The estimator in (11) is evaluated using Sig and RSig at increasing truncation levels, as well as Sup, which uses the conventional supremum distance $\rho_{\sup}(x,y) = \sup_t |x_t - y_t|$. Although both signature-based distances visibly outperform the supremum distance, we observe that Sig is sensitive to outliers, see the discussion in Section 3.3.

Finally, Table 1 illustrates the convergence studied in Theorem 3.4 and Corollary 3.11 as the number of samples M increases. In addition to Sig, RSig and Sup, we also include the methods p-var and L^p based on the metrics induced by classical L^p- and p-variation norms (see (21) in Appendix B.1). For each training sample size, the table reports the root mean-squared error (RMSE) evaluated on the independent testing set of size $M_{te} = |I_{te}| = 8192$. The last row additionally presents the maximal running time (in seconds) of the procedure, including the hyperparameter optimization and the evaluation of the RMSE for each method for tha largest training sample size $M_{tr} = 8192$.

The results clearly show that Sig and RSig substantially outperform the estimators based on conventional metrics, both in terms of accuracy and computational efficiency. Among the signature methods, RSig consis-

the log-signature, but observed no noticeable difference in performance.

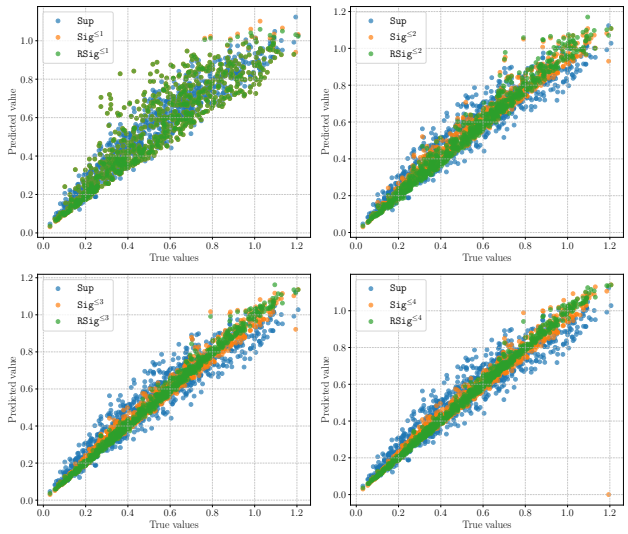

Figure 1: Scatter plots of the testing data $\{(Y^{(m)}, \widehat{Y}^{(m)}) : m \in I_{te}\}$, using signature and supremum metrics in (11). At truncation level $N = 4$, the point • near $(1.2, 0)$ illustrates an outlier of the basic method `Sig`; see the discussion in Section 3.3.

Dataset	DTWD	DTWA	SigP	Sup	L^2	RSig	Sig
ArticularyWord	**98.7**	**98.7**	97.7	92.3	94.3	97.0	95.7
AtrialFibrillation	20.0	26.7	**46.7**	33.3	33.3	26.7	33.3
BasicMotions	97.5	**100**	**100**	52.5	67.5	95.0	92.5
Cricket	**100**	**100**	95.8	58.3	81.9	83.3	75.0
Epilepsy	96.4	**97.8**	95.7	47.8	53.6	73.1	44.9
Ethanol	32.3	31.6	**43.3**	25.1	25.1	38.8	24.7
Ering	91.5	92.6	**94.8**	81.9	87.8	81.1	62.6
FaceDetection	52.9	52.8	**61.4**	50.9	52.5	51.5	51.3
FingerMovements	**53.0**	51.0	52.0	49.0	49.0	48.0	49.0
HandMovement	18.9	20.3	20.3	20.3	20.3	**29.7**	27.0
Handwriting	**60.7**	**60.7**	37.9	17.4	12.6	28.7	12.1
Heartbeat	71.7	69.3	69.8	72.2	**74.6**	71.7	72.6
Libras	87.2	88.3	**93.9**	82.8	71.1	82.8	77.8
LSST	55.1	56.7	**56.9**	10.4	16.7	41.3	31.5
NATOPS	88.3	88.3	**92.2**	75.6	76.1	81.7	76.1
PenDigits	**97.7**	**97.7**	97.4	91.4	96.3	88.4	95.1
Racketsports	80.3	84.2	**90.8**	56.6	82.2	79.6	74.3
SCP1	77.5	78.5	**78.8**	50.2	50.2	77.1	68.9
SCP2	**53.9**	52.2	50.6	50.0	51.1	**53.9**	52.2
StandWalkJump	20.0	33.3	**46.7**	**46.7**	13.3	33.3	33.3
UWaveGesture	90.3	90.0	**90.9**	82.5	84.4	83.8	80.6

Table 2: Classification accuracy (%) comparison across 21 benchmark time series datasets.

tently achieve better performance. The sensitivity of `Sig`, discussed in Section 3.3, becomes apparent as its accuracy worsens when the truncation level increases from 3 to 4, whereas `RSig` continues to improve. The price to pay lies in the additional computational effort required for the robustification. Nevertheless, it remains highly efficient—comparable to the "simpler" methods $\mathtt{L}^p$ and `Sup` —whereas p-`var` methods become impractical for this task. In Appendix C.1, the results reported in Table 1 are illustrated visually.

We conduct a second numerical experiment with synthetic data in Appendix C.2, which concerns learning path-dependent functionals of fractional Brownian motion, and we compare the performance of our nonparametric method with parametric regression on the signature.

4.2 Time Series Classification

We evaluate our signature-based Nadaraya-Watson classifier on the UEA time series classification archive (Bagnall et al., 2018)[3]. Our experimental setup includes comparisons with some distance-based time series classifiers (Abanda et al., 2019), such as dynamic time warping (`DTW`) [4], canonical signature pipeline (`SigP`) with a random forest classifier (Morrill et al., 2020) and an analysis of different distance metrics within the local regression framework.

Table 2 presents the classification accuracy (expressed as percentages) across 21 datasets. We make the following key observations. `Sup` and $\mathtt{L}^2$ serve as natural baselines for our methods, since all share the same kernel regression framework and differ only in the choice of distance. Across datasets, we observe that `Sup` and $\mathtt{L}^2$ frequently yield lower accuracy. This demonstrates that the signature distance offers a more expressive and robust similarity measure for sequential data than pointwise metrics.

`DTW`-based methods remain strong performers on some datasets (e.g., Cricket, Handwriting), reflecting their effectiveness when local temporal shifts are the dominant source of variability. However, both `Sig` and `RSig` achieve comparable or better accuracy on many datasets without requiring explicit alignment.

The computational complexity of `Sig` is of order $\mathcal{O}((M_{te} + M_{tr})Ld^N + M_{te}M_{tr}d^N)$ which is linear in sequence length L and linear in the number of training samples M_{tr}, with the exponential dependence on d^N controlled by choosing small truncation levels (typically $N \leq 5$). `DTW` methods on the other hand, scales quadratically in L, making it unpractical for a long time series. For the comparison, on the Ethanol dataset, which has the longest sequence length among our benchmark datasets, `DTW`-based classification re-

[3] We follow the dataset subset used in the original work Bagnall et al. (2018), rather than selecting a subset ourselves. This ensures consistency and allows for a reproducible comparison with prior results.

[4] `DTWD`/`DTWA` refer to 1-NN with `DTW` distance using uniform and distance-weighted voting respectively (Abanda et al., 2019)

quires approximately 1018.49 seconds for complete evaluation (on CPU). In comparison, our signature-based method with fixed bandwidth and truncation level $N = 5$ completes in just 4.89 seconds (also on CPU, without GPU acceleration).

5 Conclusion

In this paper, we have proposed a signature-based Nadaraya-Watson estimator for nonparametric regression and classification on path spaces. We provide a rigorous finite-sample guarantee.

Experimental validation on SDE learning and time series classification demonstrates consistent improvements over conventional distance metrics. While signature-based methods may not always outperform specialized techniques like DTW in specific domains, they provide a unified framework that works well across diverse sequential data types without requiring domain-specific preprocessing.

Several directions for future work can be considered. On the theoretical side, it would be interesting to extend the analysis to higher orders of smoothness on path space, in order to obtain finer approximation results and a better understanding of the underlying regularity. On the numerical side, one may investigate generalized signature distances based on alternative augmentations, such as lead–lag transforms or rectilinear interpolation; see Chevyrev and Kormilitzin (2026) for details. In a similar vein, another promising direction is to look into distances induced by "weighted" signatures recently considered in the literature Abi Jaber and Sotnikov (2025); Hager et al. (2026); Bloch et al. (2026), and to study the extent to which these richer representations improve the modeling of temporal dependence in non-Markovian time series. Finally, another natural direction is to develop regression estimators at the level of probability laws, building on expected signature distances Lemercier et al. (2021b); Chevyrev and Oberhauser (2022); Lucchese et al. (2025), with the aim of learning law-dependent functionals from path-valued data.

Acknowledgements

CB and LP gratefully acknowledge funding by Deutsche Forschungsgemeinschaft through SFB TRR 388 Project B03. The work of DG was supported by the LMBayes project (Linguistic Meaning and Bayesian Modelling).

References

Amaia Abanda, Usue Mori, and Jose A Lozano. A review on distance based time series classification. *Data Mining and Knowledge Discovery*, 33(2):378–412, 2019.

Eduardo Abi Jaber and Dimitri Sotnikov. Exponentially fading memory signature. *arXiv preprint arXiv:2507.03700*, 2025.

Germán Aneiros, Silvia Novo, and Philippe Vieu. Variable selection in functional regression models: A review. *Journal of Multivariate Analysis*, 188:104871, 2022.

Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The UEA multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*, 2018.

Fabrice Baudoin, Qi Feng, and Cheng Ouyang. Density of the signature process of fBm. *Transactions of the American Mathematical Society*, 373(12):8583–8610, 2020.

Christian Bayer, Luca Pelizzari, and John Schoenmakers. Primal and dual optimal stopping with signatures. *Finance and Stochastics*, pages 1–34, 2025a.

Christian Bayer, Luca Pelizzari, and Jia-Jie Zhu. Pricing american options under rough volatility using deep-signatures and signature-kernels. *arXiv preprint arXiv:2501.06758*, 2025b.

Linus Bleistein, Adeline Fermanian, Anne-Sophie Jannot, and Agathe Guilloux. Learning the dynamics of sparsely observed interacting systems. In *International Conference on Machine Learning*, pages 2603–2640. PMLR, 2023.

Alexandre Bloch, Samuel N Cohen, Terry Lyons, Joël Mouterde, and Benjamin Walker. The exponentially weighted signature. *arXiv preprint arXiv:2603.19198*, 2026.

Horatio Boedihardjo, Xi Geng, Terry Lyons, and Danyu Yang. The signature of a rough path: uniqueness. *Advances in Mathematics*, 293:720–737, 2016.

Jean-Philippe Bouchaud, Julius Bonart, Jonathan Donier, and Martin Gould. *Trades, quotes and prices: financial markets under the microscope.* Cambridge University Press, 2018.

Thomas Cass and Cristopher Salvi. Lecture notes on rough paths and applications to machine learning. *arXiv preprint arXiv:2404.06583*, 2024.

Kuo-Tsai Chen. Integration of paths, geometric invariants and a generalized Baker-Hausdorff formula. *Annals of Mathematics*, 65(1):163–178, 1957.

Ilya Chevyrev and Andrey Kormilitzin. A primer on the signature method in machine learning. *Signature Methods in Finance*, page 3, 2026.

Ilya Chevyrev and Harald Oberhauser. Signature moments to characterize laws of stochastic processes. *The Journal of Machine Learning Research*, 23(1): 7928–7969, 2022.

Samuel N Cohen, Silvia Lui, Will Malpass, Giulia Mantoan, Lars Nesheim, Aureo de Paula, Andrew Reeves, Craig Scott, Emma Small, and Lingyi Yang. Nowcasting with signature methods. *arXiv preprint arXiv:2305.10256*, 2023.

Adeline Fermanian. Embedding and learning with signatures. *Computational Statistics & Data Analysis*, 157:107148, 2021.

Adeline Fermanian. Functional linear regression with truncated signatures. *Journal of Multivariate Analysis*, 192:105031, 2022.

Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis: theory and practice*. Springer, 2006.

B Folland. Modern techniques and their applications. *Real Analysis (Pure and Applied Mathematics)*, 1999.

Peter K Friz and Paul P Hager. Expected signature kernels for Lévy rough paths. *arXiv preprint arXiv:2509.07893*, 2025.

Peter K Friz and Nicolas B Victoir. *Multidimensional stochastic processes as rough paths: theory and applications*, volume 120. Cambridge University Press, 2010.

Alex Graves, Marcus Liwicki, Horst Bunke, Jürgen Schmidhuber, and Santiago Fernández. Unconstrained on-line handwriting recognition with recurrent neural networks. *Advances in neural information processing systems*, 20, 2007.

Haotian Gu, Xin Guo, Timothy L Jacobs, Philip Kaminsky, and Xinyu Li. Transportation marketplace rate forecast using signature transform. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4997–5005, 2024.

Xin Guo, Binnan Wang, Ruixun Zhang, and Chaoyi Zhao. On consistency of signature using Lasso. *Operations Research*, 2025.

Paul P Hager, Fabian N Harang, Luca Pelizzari, and Samy Tindel. The Volterra signature. *arXiv preprint arXiv:2603.04525*, 2026.

Ben Hambly and Terry Lyons. Uniqueness for the signature of a path of bounded variation and the reduced path group. *Annals of Mathematics*, pages 109–167, 2010.

Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65–69, 2019.

Blanka Horvath, Maud Lemercier, Chong Liu, Terry Lyons, and Cristopher Salvi. Optimal stopping via distribution regression: a higher rank signature approach. *arXiv preprint arXiv:2304.01479*, 2023.

Patrick Kidger, Patric Bonnier, Imanol Perez Arribas, Cristopher Salvi, and Terry Lyons. Deep signature transforms. *Advances in Neural Information Processing Systems*, 32, 2019.

Franz J Király and Harald Oberhauser. Kernels for sequentially ordered data. *Journal of Machine Learning Research*, 20(31):1–45, 2019.

Anthony Knapp. *Lie groups beyond an introduction*, volume 140. Springer, 1996.

Shigeo Kusuoka and Daniel Stroock. Applications of the Malliavin calculus, part iii. *J. Fac. Sci. Univ. Tokyo Sect IA Math*, 34:391–442, 1987.

Darrick Lee and Harald Oberhauser. The signature kernel. In *Signature Methods in Finance: An Introduction with Computational Applications*, pages 85–124. Springer, 2025.

Maud Lemercier, Cristopher Salvi, Thomas Cass, Edwin V Bonilla, Theodoros Damoulas, and Terry J Lyons. SigGPDE: Scaling sparse gaussian processes on sequential data. In *International Conference on Machine Learning*, pages 6233–6242. PMLR, 2021a.

Maud Lemercier, Cristopher Salvi, Theodoros Damoulas, Edwin Bonilla, and Terry Lyons. Distribution regression for sequential data. In *International Conference on Artificial Intelligence and Statistics*, pages 3754–3762. PMLR, 2021b.

Maud Lemercier, Terry Lyons, and Cristopher Salvi. Log-PDE methods for rough signature kernels. *arXiv preprint arXiv:2404.02926*, 2024.

Wenbo V Li and Q-M Shao. Gaussian processes: inequalities, small ball probabilities and applications. *Handbook of Statistics*, 19:533–597, 2001.

Heng Lian. Convergence of nonparametric functional regression estimates with functional responses. *Electronic journal of statistics*, 2012.

Lorenzo Lucchese, Mikko S Pakkanen, and Almut ED Veraart. Learning with expected signatures: Theory and applications. In *International Conference on Machine Learning*, pages 40995–41055. PMLR, 2025.

Terry Lyons and Andrew D McLeod. Signature methods in machine learning. *arXiv preprint arXiv:2206.14674*, 2022.

Terry Lyons and Nicolas Victoir. An extension theorem to rough paths. *Annales de l'IHP Analyse non linéaire*, 24(5):835–847, 2007.

Terry J Lyons. Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, 14(2): 215–310, 1998.

Elias Masry. Nonparametric regression estimation for dependent functional data: asymptotic normality. *Stochastic Processes and their Applications*, 115 (1):155–177, 2005.

Alexander Meister. Optimal classification and nonparametric regression for functional data. *Bernoulli*, 22(3):1729 – 1744, 2016. doi: 10.3150/15-BEJ709.

Fernando Moreno-Pino, Álvaro Arroyo, Harrison Waldon, Xiaowen Dong, and Álvaro Cartea. Rough transformers: Lightweight and continuous time series modelling through signature patching. *Advances in Neural Information Processing Systems*, 37:106264–106294, 2024.

Sam Morley and Terry Lyons. Roughpy: streaming data is rarely smooth. *Proceedings of the 23rd Python in Science Conference*, 2024.

James Morrill, Adeline Fermanian, Patrick Kidger, and Terry Lyons. A generalised signature method for multivariate time series feature extraction. *arXiv preprint arXiv:2006.00873*, 2020.

Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

Philip E Protter. *Stochastic Integration and Differential Equations*. Springer, 2005.

Rimhak Ree. Lie elements and an algebra associated with shuffles. *Annals of Mathematics*, 68(2):210–220, 1958.

Jeremy F. Reizenstein and Benjamin Graham. Algorithm 1004: The iisignature library: Efficient calculation of iterated-integral signatures and log signatures. *ACM Trans. Math. Softw.*, 46(1), March 2020. ISSN 0098-3500.

Christophe Reutenauer. Free Lie algebras. In *Handbook of algebra*, volume 3, pages 887–903. Elsevier, 2003.

Cristopher Salvi, Thomas Cass, James Foster, Terry Lyons, and Weixin Yang. The signature kernel is the solution of a Goursat PDE. *SIAM Journal on Mathematics of Data Science*, 3(3):873–899, 2021.

Alexander Schell and Rima Alaifari. Nonparametric regression of stochastic processes via signatures. *Opt. Express*, 31(5):9052–9071, 2023.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

Leonie Selk and Jan Gertheiss. Nonparametric regression and classification with functional, categorical, and mixed covariates. *Advances in Data Analysis and Classification*, 17(2):519–543, 2023.

Han Lin Shang. Bayesian bandwidth estimation for a functional nonparametric regression model with mixed types of regressors and unknown error density. *Journal of Nonparametric Statistics*, 26(3):599–615, 2014.

Charles J. Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360, 1980.

Csaba Toth and Harald Oberhauser. Bayesian learning from sequential data using gaussian processes with signature covariances. In *International Conference on Machine Learning*, pages 9548–9560. PMLR, 2020.

Csaba Tóth, Harald Oberhauser, and Zoltán Szabó. Random fourier signature features. *SIAM Journal on Mathematics of Data Science*, 7(1):329–354, 2025.

Benjamin Walker, Andrew D McLeod, Tiexin Qin, Yichuan Cheng, Haoliang Li, and Terry Lyons. Log neural controlled differential equations: the lie brackets make a difference. In *Proceedings of the 41st International Conference on Machine Learning*, pages 49822–49844, 2024.

Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

Weixin Yang, Terry Lyons, Hao Ni, Cordelia Schmid, and Lianwen Jin. Developing the path signature

methodology and its application to landmark-based human action recognition. In *Stochastic Analysis, Filtering, and Stochastic Optimization: A Commemorative Volume to Honor Mark HA Davis's Contributions*, pages 431–464. Springer, 2022.

Laurence C Young. An inequality of the Hölder type, connected with Stieltjes integration. *Acta Mathematica*, 1936.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. No

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes
   (b) Complete proofs of all theoretical results. Yes
   (c) Clear explanations of any assumptions. Yes

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). No
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). No

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes
   (b) The license information of the assets, if applicable. No
   (c) New assets either in the supplemental material or as a URL, if applicable. No
   (d) Information about consent from data providers/curators. Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

Supplementary Materials

A Convergence Guarantee

In this section, we provide the detailed proof of Theorem 3.4. We begin by stating Bernstein's inequality, which serves as our main probabilistic tool.

Theorem A.1 (Bernstein's Inequality) *Let $\eta_1, \eta_2, \ldots, \eta_n$ be independent random variables that satisfy the moment condition*

$$\mathbb{E}[|\eta_i - \mathbb{E}[\eta_i]|^k] \leq \frac{1}{2} k! L^{k-2} \sigma^2, \quad \forall k \geq 2, \tag{12}$$

for some positive $L > 0$ and σ. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} \eta_i - \mathbb{E}[\eta_i] \right| \leq \sqrt{\frac{2\sigma^2 \log(2/\delta)}{n}} + \frac{L \log(2/\delta)}{n}.$$

Moment condition holds, in particular, for bounded random variables with bounded variance

$$|\eta_i| \leq \frac{L}{2} a.s, \quad \mathbb{E}[\eta_i^2] \leq \sigma^2. \tag{13}$$

The proof strategy follows a standard bias-variance decomposition approach: we first decompose the pointwise risk into bias and variance components, then we apply concentration inequalities to bound the variance terms with high probability.

Let us fix $h > 0$ and consider the estimator (3) when the number of observations M goes to infinity

$$F_h(x) = \frac{\int y K(h^{-1} \varrho(x, z)) dP(y, z)}{\int K(h^{-1} \varrho(x, z)) dP_X(z)}.$$

We also need the following notations

$$p_M(x) = \frac{1}{M} \sum_{j=1}^{M} K(h^{-1} \varrho(x, X^{(j)})), \quad \text{and} \quad p_h(x) = \int K(h^{-1} \varrho(x, z)) dP_X(z).$$

By simple algebraic manipulations, we have

$$\begin{aligned}\left|\widehat{F}(x)-F(x)\right| &\le \left|\widehat{F}(x)-F_h(x)\right| + |F_h(x)-F(x)| \\ &= \left|\frac{1}{p_M(x)}\left(\frac{1}{M}\sum_{i=1}^{M} Y^{(i)} K(h^{-1}\varrho(x,X_i)) - p_M(x)F_h(x)\right)\right| + |F_h(x)-F(x)| \\ &\le \underbrace{\left|1-\frac{p_h(x)-p_M(x)}{p_h(x)}\right|^{-1}}_{A_1}\left(\underbrace{\left|\frac{1}{M}\sum_{i=1}^{M} Y^{(i)}\frac{K(h^{-1}\varrho(x,X_i))}{p_h(x)} - F_h(x)\right|}_{A_2}\right. \\ &\quad \left. + \underbrace{\left|\frac{p_h(x)-p_M(x)}{p_h(x)}F_h(x)\right|}_{A_3}\right) + \underbrace{|F_h(x)-F(x)|}_{B}.\end{aligned}$$

Bound on B. We have

$$\begin{aligned}|F_h(x)-F(x)| &= \left|\frac{1}{p_h(x)}\int yK(h^{-1}\varrho(x,z))dP(y,z) - F(x)\right| \\ &= \left|\frac{1}{p_h(x)}\int F(z)K(h^{-1}\varrho(x,z))dP_X(z) - F(x)\right| \\ &\le \frac{1}{p_h(x)}\left(\int K(h^{-1}\varrho(x,z))|F(z)-F(x)|dP_X(z)\right).\end{aligned}$$

Applying conditions (4) and (6) yields

$$|F_h(x)-F(x)| \le \frac{1}{p_h(x)}\int K(h^{-1}\varrho(x,z))\varrho(x,z)^{\beta}dP_X(z) \le \frac{h^{\beta}B\phi_x(h)}{p_h(x)}.$$

Note that under the condition 6, we have

$$p_h(x) = \int K(h^{-1}\varrho(x,z))dP_X(z) \ge b\mathbb{P}[\varrho(X,x)\le h] = b\phi_x(h). \tag{14}$$

This leads to the bound on the bias term

$$|F_h(x)-F(x)| \le \frac{Bh^{\beta}}{b}. \tag{15}$$

Bound on A_1. Provided that $|\frac{p_h(x)-p_M(x)}{p_h(x)}| < 1/2$ we have

$$\left|1-\frac{p_h(x)-p_M(x)}{p_h(x)}\right|^{-1} \le 2.$$

So it remains to show that $|\frac{p_h(x)-p_M(x)}{p_h(x)}| < 1/2$, which is given by the following proposition.

Proposition A.2 *Let*

$$M \ge \frac{16B\log(6/\delta)}{b\phi_x(h)}. \tag{16}$$

Then, for $\delta \in (0,1)$, with probability at least $1-\delta/3$

$$\frac{|p_h(x)-p_M(x)|}{p_h(x)} \le \frac{1}{2}. \tag{17}$$

Proof To establish the result, we verify the conditions (13) of Bernstein's inequality for the random variables

$$\eta_i = \frac{K(h^{-1}\varrho(x, X^{(i)}))}{p_h(x)}.$$

Since the kernel is bounded, we have

$$|\eta_i| \leq \frac{B}{p_h(x)}.$$

For the variance,

$$\mathbb{E}[\eta_i^2] = \int \frac{K(h^{-1}\varrho(x,z))^2}{p_h^2(x)} dP_X(z) \leq \frac{B}{p_h(x)} \leq \frac{B}{b\phi_x(h)},$$

where the last inequality follows from (14). Applying Bernstein's inequality, we conclude that with probability at least $1 - \delta/3$

$$\frac{|p_h(x) - p_M(x)|}{p_h(x)} \leq \sqrt{\frac{2B\log(6/\delta)}{b\phi_x(h)M}} + \frac{2B\log(6/\delta)}{b\phi_x(h)M}. \tag{18}$$

The proposition follows from (16). □

Bound on A_3. Since $|Y| \leq R$, we have $|F_h(x)| \leq R$. Therefore, using (18) we have

$$\left|\frac{p_h(x) - p_M(x)}{p_h(x)} F_h(x)\right| \leq R\left(\sqrt{\frac{2B\log(6/\delta)}{b\phi_x(h)M}} + \frac{2B\log(6/\delta)}{b\phi_x(h)M}\right). \tag{19}$$

Bound on A_2. The bound follows from the following

Lemma A.3 *Let $|Y| \leq R$. Then, with probability at least $1 - \delta/3$*

$$\left|\frac{1}{M}\sum_{i=1}^{M} Y^{(i)} \frac{K(h^{-1}\varrho(x, X^{(i)}))}{p_h(x)} - F_h(x)\right| \leq R\sqrt{\frac{2B\log(6/\delta)}{b\phi_x(h)M}} + \frac{2RB\log(6/\delta)}{b\phi_x(h)M}. \tag{20}$$

Proof We check the Bernstein condition (13) for $\eta_i = Y^{(i)} \frac{K(h^{-1}\varrho(x,X^{(i)}))}{p_h(x)}$. We have

$$|\eta_i| \leq \frac{RB}{b\phi_x(h)}, \quad \mathbb{E}[|\eta_i|^2] \leq \frac{R^2 B}{b\phi_x(h)}.$$

The bound (20) follows from the Bernstein inequality. □

Proof [proof of Theorem 3.4] Applying the union bound, we get, with probability at least $1 - \delta$

$$\begin{aligned}|\widehat{F}(x) - F_h(x)| &\leq 2(R + |F_h(x)|)\left(\sqrt{\frac{2B\log(6/\delta)}{b\phi_x(h)M}} + \frac{2B\log(6/\delta)}{b\phi_x(h)M}\right) \\ &\leq 8R\sqrt{\frac{2B\log(6/\delta)}{b\phi_x(h)M}}.\end{aligned}$$

The last inequality, together with the bias bound (15), finishes the proof. □

B Signature appendix

In this section, we provide a supplementary introduction to path signatures, complementing Section 3, and present the proofs of the convergence rates for local signature regression. Our primary references for signatures and rough paths are the classical monograph Friz and Victoir (2010) and the recent lecture notes Cass and Salvi (2024), to which we refer for further details.

B.1 Signatures and tensor algebras

As anticipated in Section 3.1, the signature (or path signature) of a continuous path $x : [0,T] \to \mathbb{R}^d$ is the collection of *iterated integrals* against itself. To give a meaning to this object, one needs a suitable notion of regularity for paths $x : [0,T] \to \mathbb{R}^d$.

Definition B.1 *For any real number $p \geq 1$ and continuous path $x : [0,T] \to \mathbb{R}^d$, we define the p-variation by*

$$\|x\|_{p-var} = \left(\sup_{\mathcal{P} \subset [0,T]} \sum_{[u,v] \in \mathcal{P}} |x_v - x_u|^p \right)^{1/p}, \tag{21}$$

where $|\cdot|$ is the Euclidean norm on $\mathbb{R}^d$, and the supremum is taken over all partitions $\mathcal{P}$ of $[0,T]$. We denote by $C^{p-var}([0,T], \mathbb{R}^d)$ the spaces of all continuous paths of finite p-variation, that is $\|x\|_{p-var} < \infty$.

It is well known that for any $1 \leq p \leq p' < \infty$ one has the inclusion $C^{p'\text{-var}} \subset C^{p\text{-var}}$ (Friz and Victoir, 2010, Proposition 5.3). In particular, every continuously differentiable path $x \in C^1$ has finite 1-variation with

$$\|x\|_{1\text{-var}} = \int_0^T |\dot{x}_t| \, dt,$$

see (Friz and Victoir, 2010, Proposition 1.27). Increasing $p \geq 1$ enlarges the class $C^{p\text{-var}}$, admitting more irregular paths, such as $\lfloor 1/p \rfloor$-Hölder paths (Friz and Victoir, 2010, Proposition 5.2).

While the class of p-variation paths provides the correct analytical framework to define signatures, let us now turn to the algebraic aspects. For multidimensional paths x, iterated integrals naturally appear as tensors

$$\left(\int_0^T dx_t^i \right)_{i \in [d]} \in \mathbb{R}^d, \qquad \left(\int_0^T \int_0^t dx_r^i \, dx_t^j \right)_{i,j \in [d]} \in \mathbb{R}^d \otimes \mathbb{R}^d, \quad \dots$$

recalling the index notation $[d] = \{1, \dots, d\}$. More generally, the n-fold iterated integrals take values in $(\mathbb{R}^d)^{\otimes n}$.

Let $\{e_1, \dots, e_d\}$ denote the canonical basis of $\mathbb{R}^d$. For any word $w = i_1 \cdots i_n$ with letters from the alphabet $\mathcal{A} = [d]$, we denote by $e_w = e_{i_1} \otimes \cdots \otimes e_{i_n}$ the corresponding basis element of $(\mathbb{R}^d)^{\otimes n}$. Starting from the representation $x = \sum_{i=1}^d e_i x^i$, one can then write the n-fold iterated integral tensor as

$$\left(\int_0^T \int_0^{t_n} \cdots \int_0^{t_2} dx_{t_1}^{i_1} \cdots dx_{t_n}^{i_n} \right)_{i_1, \dots, i_n \in [d]} = \sum_{w = i_1 \cdots i_n} \left(\int_0^T \int_0^{t_n} \cdots \int_0^{t_2} dx_{t_1}^{i_1} \cdots dx_{t_n}^{i_n} \right) e_w.$$

The full signature of a path, and its truncation at level N, take values in the spaces

$$\mathcal{T} = \prod_{k \geq 0} (\mathbb{R}^d)^{\otimes k}, \qquad \mathcal{T}^{\leq N} = \prod_{k=0}^{N} (\mathbb{R}^d)^{\otimes k},$$

where we adopt the convention $(\mathbb{R}^d)^{\otimes 0} = \mathbb{R}$. Using the basis representation of tensors introduced above, any element $\mathbf{a} \in \mathcal{T}$ can be represented through the series

$$\mathbf{a} = \sum_{w \in \mathcal{W}} \mathbf{a}^w e_w, \quad \mathbf{a}^w \in \mathbb{R},$$

where $\mathcal{W}$ denotes the space of all words. Additionally, we denote by $\mathbf{a}^{(k)}$ the projection of $\mathbf{a}$ to the tensor-level k, that is

$$\mathbf{a}^{(k)} = \sum_{w \in \mathcal{W}^{(k)}} \mathbf{a}^w e_w, \qquad \mathcal{W}^{(k)} = \{w = i_1 \cdots i_k : i_1, \dots, i_k \in [d]\} \subset \mathcal{W}.$$

Finally, we equip $\mathcal{T}$, as well as its truncated version, with the product

$$\otimes : \mathcal{T} \times \mathcal{T} \to \mathcal{T}, \qquad \left(\sum_{w \in \mathcal{W}} \mathbf{a}^w e_w \right) \otimes \left(\sum_{w \in \mathcal{W}} \mathbf{b}^w e_w \right) = \sum_{w \in \mathcal{W}} \left(\sum_{l=0}^{|w|} \mathbf{a}^{w_1 \cdots w_l} \mathbf{b}^{w_{l+1} \cdots w_{|w|}} \right) e_w,$$

which turns $(\mathcal{T}, \otimes)$ into an algebra, often called the *extended tensor algebra*; see (Cass and Salvi, 2024, Chapter 1.1.2). Moreover, we can also naturally define addition

$$+ : \mathcal{T} \times \mathcal{T} \to \mathcal{T}, \qquad \left(\sum_{w \in \mathcal{W}} \mathbf{a}^w e_w\right) + \left(\sum_{w \in \mathcal{W}} \mathbf{b}^w e_w\right) = \sum_{w \in \mathcal{W}} (\mathbf{a}^w + \mathbf{b}^w) e_w.$$

Finally, we can endow $\mathcal{T}$ with a Hilbert space structure by using the inner product on $(\mathbb{R}^d)^{\otimes k}$ defined by

$$\langle v, w \rangle_{(\mathbb{R}^d)^{\otimes k}} = \prod_{l=1}^{k} \langle v_l, w_l \rangle_{\mathbb{R}^d}, \qquad v = v_1 \otimes \cdots \otimes v_k, \; w = w_1 \otimes \cdots \otimes w_k,$$

and set $\|v\|_{(\mathbb{R}^d)^{\otimes k}} = \sqrt{\langle v, v \rangle_{(\mathbb{R}^d)^{\otimes k}}}$. This extends to $\mathcal{T}$ via $\langle \mathbf{a}, \mathbf{b} \rangle_{\mathcal{T}} = \sum_{k \geq 0} \langle \mathbf{a}^{(k)}, \mathbf{b}^{(k)} \rangle_{(\mathbb{R}^d)^{\otimes k}}$, which induces the Hilbert space

$$\mathfrak{T} = \left\{ \mathbf{a} \in \mathcal{T} : \|\mathbf{a}\| = \sqrt{\langle \mathbf{a}, \mathbf{a} \rangle_{\mathcal{T}}} < \infty \right\} \subset \mathcal{T},$$

see (Cass and Salvi, 2024, Section 1.1.2) for further details.

We are now ready to defined the signature which dates back to Chen (1957), introduced here as a mapping from p-variation spaces into the extended tensor algebra. As in the case of continuously differentiable paths $\mathcal{X} = C^1([0,T], \mathbb{R}^d)$ discussed in Section 3, the signature can immediately be define for p-variation paths with $1 \leq p < 2$. This is made possible by Young's generalization of the Riemann–Stieltjes integral Young (1936); for further background we refer to (Friz and Victoir, 2010, Chapter 6).

Definition B.2 *For any real number $1 \leq p < 2$ and word $w = i_1 \cdots i_n \in \mathcal{W}$, we define*

$$\mathrm{Sig}(\cdot)^w : C^{p-var}([0,T], \mathbb{R}^d) \to \mathbb{R}, \quad x \mapsto \mathrm{Sig}(x)^w = \int_0^T \int_0^{t_n} \cdots \int_0^{t_2} dx_{t_1}^{i_1} \cdots dx_{t_k}^{i_k}, \tag{22}$$

and the k-th signature level is then by $\mathrm{Sig}(x)^{(k)} = \sum_{w \in \mathcal{W}^{(k)}} \mathrm{Sig}(x)^w e_w$. Finally, the full signature is defined as

$$\mathrm{Sig} : C^{p-var}([0,T], \mathbb{R}^d) \to \mathcal{T}, \quad x \mapsto \mathrm{Sig}(x) = \sum_{w \in \mathcal{W}} \mathrm{Sig}(x)^w e_w. \tag{23}$$

In the case of p-variation paths with $p > 2$, it is no longer clear whether (22) is well defined, and more sophisticated constructions are required to introduce the *rough path signature*. To keep this introduction concise, we only state the following remark and refer the interested reader to the cited references for details.

Remark B.3 *For more irregular paths, such as Brownian sample paths where $p > 2$, Young's extension of the Riemann–Stieltjes integral is no longer sufficient to define the signature. As already observed by Young (1936), $\alpha > 1/2$ (equivalently $p < 2$) is necessary and sufficient for a well-defined notion of the integral for Hölder paths. One of the major achievements of Lyons' rough path theory Lyons (1998) was to realize that the first $\lfloor p \rfloor$ signature levels of x contain exactly the missing information needed to extend integration to irregular paths. By enhancing x to a* rough path

$$x \rightsquigarrow \mathbf{x} = \left(\mathrm{Sig}(x)^{(1)}, \ldots, \mathrm{Sig}(x)^{(\lfloor p \rfloor)}\right),$$

where these abstract components mimic the classical signature, a consistent integration theory can be developed with respect to $\mathbf{x}$. This provides a robust and deterministic framework for differential equations driven by Brownian motion and more general stochastic processes. In particular, Lyons' extension theorem (Lyons, 1998, Theorem 2.2.1) ensures that the signature of a rough path $\mathbf{x}$ is again well defined and enjoys the same algebraic properties as in Definition B.2. Moreover, for a large class of stochastic processes the lift $x \mapsto \mathbf{x}$ is well understood; for instance, semimartingales can be lifted using Itô calculus. We refer to Lyons and Victoir (2007); Friz and Victoir (2010); Cass and Salvi (2024) for further details.

One of the most fundamental properties of the signature, which we rely on multiple times in this article, is the fast decay of $\mathrm{Sig}(x)^{(k)}$ as k increases. For the case $p = 1$, the following lemma is an elementary exercise; see (Cass and Salvi, 2024, Proposition 1.2.3), and for more irregular paths see (Friz and Victoir, 2010, Section 9.1.1).

Lemma B.4 *For any $x \in C^{1\text{-}var}([0,T],\mathbb{R}^d)$ we have*

$$\|\mathrm{Sig}(x)^{(k)}\|_{(\mathbb{R}^d)^{\otimes k}} \leq \frac{\|x\|_{1\text{-}var}^k}{k!}, \qquad \forall k \in \mathbb{N}.$$

In particular, $\mathrm{Sig}(C^{1-var}) \subset \mathfrak{T}$.

The above lemma ensures, in particular, that the signature semi-distance in (8),

$$\varrho^{\mathrm{Sig}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+, \qquad \varrho^{\mathrm{Sig}}(x,y) = \|\mathrm{Sig}(x) - \mathrm{Sig}(y)\|,$$

as well as its truncated version in (9), are well defined. Indeed, an application of the triangle inequality shows that for all $x, y \in C^{1-var}$, we have

$$\varrho^{\mathrm{Sig}}(x,y)^2 := \sum_{k\geq 0} \|\mathrm{Sig}(x)^{(k)} - \mathrm{Sig}(y)^{(k)}\|^2_{(\mathbb{R}^d)^{\otimes k}} \leq C \sum_{k\geq 0} \|\mathrm{Sig}(x)^{(k)}\| + \|\mathrm{Sig}(y)^{(k)}\|_{(\mathbb{R}^d)^{\otimes k}} \leq C \exp(\|x\|_{1-var} + \|y\|_{1-var}).$$

It is important to note, however, that ϱ^{Sig} is in general not a true metric, as the following lemma demonstrates.

Lemma B.5 *Let $x \in C^{p-var}$ with $1 \leq p < 2$ and $\tau : [0,T] \to [0,T]$ a continuous, non-decreasing surjection. Then*

$$\varrho^{\mathrm{Sig}}(x, x \circ \tau) = 0.$$

This is a direct consequence of the invariance of the signature under time reparametrization; see (Friz and Victoir, 2010, Proposition 7.10). The corresponding equivalence classes of paths

$$[x] = \{y \in C^{p\text{-var}} : \varrho^{\mathrm{Sig}}(x,y) = 0\}, \qquad x \in C^{p\text{-var}},$$

are well understood and known as *tree-like equivalence*. This was established in Hambly and Lyons (2010) for paths of bounded variation and extended to $p > 1$ in Boedihardjo et al. (2016); we refer to the latter references for a precise definition.

To obtain a genuine distance, one natural approach is to shrink the space by identifying tree-like equivalent paths $x \sim y$, i.e. by working on the quotient $\mathcal{X}/_\sim$. The alternative considered in this paper is to augment all paths with time, namely

$$\mathcal{X} = \widehat{C}^{p\text{-var}}([0,T],\mathbb{R}^{d+1}) = \big\{t \mapsto (t, x_t) : x \in C^{p\text{-var}}([0,T],\mathbb{R}^d)\big\}. \tag{24}$$

Finally we note that signatures are by construction invariant with respect to the initial condition, that is $\varrho^{\mathrm{Sig}}(x, x+c) = 0$, so that ϱ_{Sig} only defines a true metric for paths with identical initial condition.

Lemma B.6 *For any $1 \leq p < 2$ and $\xi \in \mathbb{R}^d$, ϱ_{Sig} defines a true metric on $\mathcal{X}_\xi = \widehat{C}^{p-var} \cap \{\widehat{x} : \widehat{x}_0 = (0,\xi)\}$.*

Proof Since $(\mathfrak{T}, \|\cdot\|)$ is a Hilbert space, symmetry and the triangle inequality follow immediately. Moreover, we have $\varrho^{\mathrm{Sig}}(x,y) \geq 0$, and therefore we are left to prove that $\varrho^{\mathrm{Sig}}(x,y) = 0 \Rightarrow x = y$. Now we can notice that for $\widehat{x}, \widehat{y} \in \widehat{C}^{p-var}$ and any word $w = i1\cdots 1$ it follows by the Cauchy formula for repeated integration that

$$\mathrm{Sig}(\widehat{x})^w = \int_0^T \int_0^{t_n} \cdots \int_0^{t_3} (x^i_{t_2} - x^i_0) dt_2 \cdots dt_n = \int_0^T (x^i_t - x^i_0) \frac{(T-t)^{n-1}}{(n-1)!} dt,$$

where n is the number of 1 in w, and $i \in [d]$. But then in particular, for all $i \in [d]$ we have

$$\varrho^{\mathrm{Sig}}(\widehat{x},\widehat{y}) = 0 \Leftrightarrow \mathrm{Sig}(\widehat{x}) - \mathrm{Sig}(\widehat{y}) = 0 \quad \Longrightarrow \quad \int_0^T (x^i_t - y^i_t)(T-t)^n dt - \frac{(y^i_0 - x^i_0)T^n}{n!} = 0,$$

for all $n \geq 0$. Since $x_0 = y_0$ on $\mathcal{X}_\xi$ and since monomials are dense in L^p, the latter is only possible if $x^i = y^i$ a.e. for all $i \in [d]$, and thus, in particular, $\widehat{x} = \widehat{y}$ almost everywhere. □

From an algebraic perspective, one of the key features of the signature is the *shuffle identity*, which plays an important role in our theoretical results. To this end, it is useful to introduce the notation of pairing words in $\mathcal{W}$ with elements in $\mathcal{T}$

$$\langle \cdot, \cdot \rangle : \mathcal{W} \times \mathcal{T} \to \mathbb{R}, \quad (w, \mathbf{a}) \mapsto \langle w, \mathbf{a} \rangle = \mathbf{a}^w,$$

which extends linearly to the span of words in the alphabet $\mathcal{A}$, that is, to the free associative algebra $\mathbb{R}\langle \mathcal{A} \rangle$. The shuffle identity is a generalization of integration by parts to higher order iterated integrals appearing in the signature. Starting with level 2 and assuming $x_0 = 0$ for simplicity, integration by parts suggests that

$$\int_0^T x_t^i dx_t^j + \int_0^T x_t^j dx_t^i = x_t^i x_t^j \quad \text{that is,} \quad \langle ij + ji, \mathrm{Sig}(x) \rangle = \langle i, \mathrm{Sig}(x) \rangle \langle j, \mathrm{Sig}(x) \rangle.$$

To capture this relation on higher levels of the signature, we introduce the shuffle-product on the space of words recursively by

$$w \sqcup\!\sqcup \emptyset = \emptyset \sqcup\!\sqcup w = w, \qquad wi \sqcup\!\sqcup vj = (w \sqcup\!\sqcup vj)i + (wi \sqcup\!\sqcup v)j,$$

which bi-linearly extends to the span of words $\mathbb{R}\langle \mathcal{A} \rangle$, so that $\sqcup\!\sqcup : \mathbb{R}\langle \mathcal{A} \rangle \times \mathbb{R}\langle \mathcal{A} \rangle \to \mathbb{R}\langle \mathcal{A} \rangle$. The integration by parts identity above then simply reads $\langle i \sqcup\!\sqcup j, \mathrm{Sig}(x) \rangle = \langle i, \mathrm{Sig}(x) \rangle \langle j, \mathrm{Sig}(x) \rangle$. Perhaps surprisingly, and first observed already in Ree (1958), is that this relation holds for arbitrary linear combinations of words

$$\langle w \sqcup\!\sqcup v, \mathrm{Sig}(x) \rangle = \langle w, \mathrm{Sig}(x) \rangle \langle v, \mathrm{Sig}(x) \rangle, \quad \forall w, v \in \mathbb{R}\langle \mathcal{A} \rangle.$$

In particular, the signature, resp. truncations thereof, take values in the following subspaces of $\mathcal{T}$

$$\mathcal{G} = \{ \mathbf{a} \in \mathcal{T} \setminus \{\mathbf{0}\} : \langle w \sqcup\!\sqcup v, \mathbf{a} \rangle = \langle w, \mathbf{a} \rangle \langle v, \mathbf{a} \rangle, \ \forall w, v \in \mathbb{R}\langle \mathcal{A} \rangle \}, \quad \mathcal{G}^{\leq N} = \{ \mathbf{g} \in \mathcal{G} : \mathbf{g}^n = 0, \ \forall n > N \},$$

which are often called *group-like elements*. It can for instance be found in (Friz and Victoir, 2010, Section 7.3.1) that $\mathcal{G}^{\leq N}$ is a Lie group associated to the free Lie algebra $\mathfrak{g}^{\leq N} \in \mathcal{T}^{\leq N}$ with bracket given by $[\mathbf{a}, \mathbf{b}] = \mathbf{a} \otimes \mathbf{b} - \mathbf{b} \otimes \mathbf{a}$, with exponential and logarithmic maps given by

$$\exp_\otimes : \mathfrak{g} \to \mathcal{G}, \qquad \mathbf{g} \mapsto \exp_\otimes(\mathbf{g}) = \sum_{n \geq 0} \frac{\mathbf{g}^{\otimes n}}{n!}, \qquad \log_\otimes : \mathcal{G} \to \mathfrak{g}, \quad \mathbf{g} \mapsto \sum_{n \geq 1} (-1)^{n+1} \frac{\mathbf{g}^{\otimes n}}{n!}$$

For a more general introduction to free Lie algebras we refer to Reutenauer (2003).

In Section 3, we made the assumption that the stochastic process $X \in \mathcal{X}$ has the property that its truncated signature $\mathrm{Sig}(X)^{\leq N} \in \mathcal{G}^{\leq N}$ admits a density with respect to the Haar measure, which we now define.

Definition B.7 *We denote by m^N (resp. μ^N) the unique*[5] *left- and right-invariant Haar measure*[6] *on the Lie algebra $\mathfrak{g}^{\leq N}$ (resp. the Lie group $\mathcal{G}^{\leq N}$).*

An important observation is that μ^N is determined by m^N through the logarithmic map, we refer to (Friz and Victoir, 2010, Proposition 16.40) for a proof.

Lemma B.8 *The Haar measure m^N coincides with the Lebesgue measure on $\mathfrak{g}^{\leq N}$, and μ^N on the Lie group $\mathcal{G}^{\leq N}$ is given by the push-forward*

$$\mu^N(A) = m^N(\log_\otimes(A)), \quad \forall A \in \mathcal{B}_{\mathcal{G}^{\leq N}}.$$

We conclude this introduction with the construction of the robust signature, introduced in Chevyrev and Oberhauser (2022) and frequently used in this work. The unbounded nature of the classical signature leads to several theoretical and practical difficulties, as outlined in the main body of this article. To address this issue, Chevyrev and Oberhauser proposed to *bound* the signature map via a tensor normalization

$$\Lambda : \mathcal{T}_1 \to \{ \mathbf{a} \in \mathcal{T}_1 : \|\mathbf{a}\| \leq R \}, \qquad R > 0, \tag{25}$$

[5] which exists and is unique up to constant factors on any locally compact group, see, e.g., (Folland, 1999, Theorem 11.8).

[6] that is, a regular Borel measure m, such that $m(gH) = m(H) = m(Hg)$ for all group elements g and Borel measurable sets H

where $\mathcal{T}_1 = \{\mathbf{a} \in \mathcal{T} : \mathbf{a}^{\emptyset} = 1\}$, and defined the robust signature as the composition $\Lambda \circ \mathrm{Sig}$. The main challenge in this construction is to ensure that the resulting feature map $x \mapsto \Lambda \circ \mathrm{Sig}(x)$ retains the key advantages of classical signatures, such as their expressivity. A natural way to construct such a normalization is through dilations,

$$\delta_\lambda : \mathcal{T}_1 \to \mathcal{T}_1, \qquad \mathbf{a} \mapsto \delta_\lambda(\mathbf{a}) = \sum_{w \in \mathcal{W}} \lambda^{|w|} \mathbf{a}^w e_w, \quad \lambda \in \mathbb{R}_+,$$

where $|w|$ defines the length of the word, that is $|i_1 \cdots i_n| = n$. Of course, setting $\Lambda(\mathbf{a}) = \delta_\lambda(\mathbf{a})$ for some fixed $\lambda > 0$ does not suffice to bound the signature map (Definition B.2) via $\Lambda \circ \mathrm{Sig}$, since one easily verifies that

$$\Lambda \circ \mathrm{Sig}(\lambda^{-1} x) = \mathrm{Sig}(x), \qquad \forall x \in C^{p\text{-var}}.$$

Hence, the parameter λ must depend on the element itself, i.e. $\Lambda(\mathbf{a}) = \delta_{\lambda(\mathbf{a})}(\mathbf{a})$, which leads to the following definition; see (Chevyrev and Oberhauser, 2022, Section 3.2).

Definition B.9 *Let $R > 0$ be fixed and $\lambda : \mathcal{T}_1 \to \mathbb{R}_+$ a function. The map*

$$\Lambda(\mathbf{a}) = \delta_{\lambda(\mathbf{a})}(\mathbf{a})$$

is called a tensor normalization *if it is continuous*[7] *and injective, and if $\|\Lambda(\mathbf{a})\| \le R$ for all $\mathbf{a} \in \mathcal{T}_1$. In this case, for any $1 \le p < 2$, the* robust signature *is defined by*

$$\mathrm{RSig} : C^{p\text{-}var}([0,T], \mathbb{R}^d) \to \mathcal{T}_1, \qquad x \mapsto \mathrm{RSig}(x) = \Lambda(\mathrm{Sig}(x)).$$

For our purposes, the most important theoretical property is that the robust signature remains injective on the space $\widehat{C}^{p\text{-var}}$ defined earlier, and in particular

$$\varrho_{\mathrm{RSig}}(x, y) = 0 \iff x = y, \qquad \forall x, y \in \widehat{C}^{p\text{-var}},$$

where $\varrho_{\mathrm{RSig}}(x,y) = \|\mathrm{RSig}(x) - \mathrm{RSig}(y)\|$. In all our numerical experiments we construct λ as proposed in (Chevyrev and Oberhauser, 2022, Example 4), which we shall briefly outline now.

Example 2 Define the mapping $\Psi = \Psi_{a,C} : [1, \infty) \to [1, \infty)$ by

$$\Psi(\sqrt{x}) = \begin{cases} x & x \le C \\ C + C^{1+a}(C^{-a} + x^{-a})/a & x > C, \end{cases}$$

for some fixed constants $a > 0$ and $C \ge 1$. Now for any $\mathbf{a} \in \mathcal{T}_1$, we define $\lambda(\mathbf{a})$ to be unique non-negative number such that

$$\|\delta_{\lambda(\mathbf{a})}(\mathbf{a})\|^2 = \sum_{k \ge 0} \lambda(\mathbf{a})^{2k} \|\mathbf{a}^{(k)}\|_{(\mathbb{R}^d)^{\otimes k}} = \psi(\|\mathbf{a}\|).$$

The resulting $\Lambda = \delta_{\lambda(\cdot)}(\cdot)$ defines a tensor-normalization.

B.2 Proofs Section 3.2

Proof [of Lemma 3.7] First we can notice that for any $x \in \mathcal{X}$, we have

$$\varrho^{\mathrm{Sig}}(X,x)^2 = \|\mathrm{Sig}(X) - \mathrm{Sig}(x)\|^2 = \varrho^{\mathrm{Sig}}_{\le N}(X,x)^2 + \sum_{k > N} \|\mathrm{Sig}(X)^{(k)} - \mathrm{Sig}(x)^{(k)}\|^2_{(\mathbb{R}^d)^{\otimes k}} \le \varrho^{\mathrm{Sig}}_{\le N}(X,x)^2 \quad \text{a.s.},$$

so that in fact globally it holds that

$$\mathbb{P}[\varrho^{\mathrm{Sig}}_{\le N}(X,x) \le h] \ge \mathbb{P}[\varrho^{\mathrm{Sig}}(X,x) \le h], \quad \forall x \in \mathcal{X}.$$

For the second part, it follows directly by definition, see also (Friz and Victoir, 2010, Proposition 7.8), that $\mathrm{Sig}(x)$ corresponds to the terminal value to the $\mathcal{T}$-valued ODE

$$\mathrm{Sig}(x)_0 = \mathbf{1} \in \mathcal{T}, \quad d\mathrm{Sig}(x)_t = \mathrm{Sig}(x)_t \otimes dx_t, \quad 0 < t \le T,$$

[7] With respect to the Banach space topology discussed at the beginning of this chapter.

where $\mathbf{1}^{\emptyset} = 1$ and $\mathbf{1}^{w} = 0$ for all $w \neq \emptyset$. Following the techniques used in Friz and Hager (2025) for *free developments* in $\mathcal{T}$, it follows from the triangle inequality that for any $x, y \in \mathcal{X}$

$$\begin{aligned}\|\mathrm{Sig}(x) - \mathrm{Sig}(y)\| &\leq \int_0^T \|\mathrm{Sig}(x)_t \otimes \dot{x}_t - \mathrm{Sig}(y)_t \otimes \dot{y}_t\| dt \\ &\leq \int_0^T \|(\mathrm{Sig}(x)_t - \mathrm{Sig}(y)_t) \otimes \dot{x}_t\| dt + \int_0^T \mathrm{Sig}(y)_t \otimes |\dot{x}_t - \dot{y}_t| dt \\ &\leq \int_0^T \|\mathrm{Sig}(x)_t - \mathrm{Sig}(y)_t\| |\dot{x}_t| dt + \int_0^t \|\mathrm{Sig}(y)_t\| |\dot{x}_t - \dot{y}_t| dt.\end{aligned}$$

An application of Lemma B.4 shows that $\int_0^T \|\mathrm{Sig}(y)_t\| |\dot{x}_t - \dot{y}_t| dt \leq e^{\|y\|_{1-var}} \|x - y\|_{1-var} =: \alpha(T)$. Applying Grönwalls inequality together with $\beta(t) = |\dot{x}_t|$, it follows that

$$\begin{aligned}\|\mathrm{Sig}(x) - \mathrm{Sig}(y)\| \leq & e^{\|y\|_{1-var}} \|x - y\|_{1-var} + \int_0^T e^{\|y\|_{1-var;[0,t]}} \|x - y\|_{1-var;[0,t]} |\dot{x}_t| e^{\|x\|_{1-var;[t,T]}} dt \\ \leq & \|x - y\|_{1-var} \left(e^{\|y\|_{1-var}} + \|x\|_{1-var} e^{\|y\|_{1-var} + \|x\|_{1-var}} \right).\end{aligned}$$

Now setting $C_{\mathcal{R}} = e^{\mathcal{R}} + \mathcal{R} e^{2\mathcal{R}}$, for any random variable $X \in \mathcal{X}_M$ we almost surely have $\varrho^{\mathrm{Sig}}(X, x) \leq C_{\mathcal{R}} \|X - x\|_{1-var}$, and therefore

$$\mathbb{P}[\varrho^{\mathrm{Sig}}(X, x) \leq h] \geq \mathbb{P}[\|X - x\|_{1-var} \leq Ch], \quad \forall x \in \mathcal{X}_{\mathcal{R}},$$

where $C = C_{\mathcal{R}}^{-1}$. □

Proof [of Proposition 3.9] For any random variable $X \in \mathcal{X}$, such that Assumption 3.8 holds true, we have

$$\phi_x^{\mathrm{Sig},N}(h) = \mathbb{P}[\varrho_{\leq N}^{\mathrm{Sig}}(X, x) \leq h] = \int_{\mathcal{G}_{h,x}^{\leq N}} p(\mathbf{g}) d\mu^N(\mathbf{g}) \geq c\mu^N(\mathcal{G}_{h,x}^{\leq N}),$$

where $\mathcal{G}_{h,x}^{\leq N} = \{\mathbf{g} \in \mathcal{G}_1^{\leq N} : \|\mathbf{g} - \mathrm{Sig}(x)^{\leq N}\| \leq h\}$. An application of Lemma B.8 together with a change of variable for the push-forward measure shows

$$\mu^N(\mathcal{G}_{h,x}^{\leq K}) = m^N\Big(\{\mathbf{g} \in \mathfrak{g}^{\leq N} : \|\exp_{\otimes}(\mathbf{g}) - \exp_{\otimes}(\mathbf{g_0}(x))\| \leq h\} \Big), \quad \mathbf{g_0}(x) = \log_{\otimes}(\mathrm{Sig}(x)^{\leq N}),$$

where $\log_{\otimes}$ is the inverse of $\exp_{\otimes}$. Now since $\mathcal{G}^{\leq N}$ is a free nilpotent, connected and simply connected Lie group, $\exp_{\otimes}$ is a diffeomorphism (Knapp, 1996, Theorem 1.127), so that in particular

$$m^N\Big(\{\mathbf{g} \in \mathfrak{g}^{\leq N} : \|\exp_{\otimes}(\mathbf{g}) - \exp_{\otimes}(\mathbf{g_0}(x))\| \leq h\} \Big) \geq m^N\Big(\{\mathbf{g} \in \mathfrak{g}^{\leq N} : \|\mathbf{g} - \mathbf{g_0}(x)\| \leq Ch\} \Big) \sim h^{\dim(\mathfrak{g}^{\leq N})},$$

since m^N is the Lebesgue measure by Lemma B.8. Finally, it follows from (Reutenauer, 2003, Theorem 6) that $\dim(\mathfrak{g}^{\leq N})$ is given by $\nu(N)$ in Lemma 3.9. □

Proof [of Corollary 3.11] From Proposition 3.9 we know $\phi_x(h) \geq \tilde{C} h^{\nu(K)}$ for some $\tilde{C} > 0$. Since all the assumption of Theorem 3.4 hold, we know that for any $\delta > 0$, with probability at least $1 - \delta$ the following bound holds

$$|\widehat{F}(x) - F(x)| \leq \frac{B}{b} h^{\beta} + 8R\sqrt{\frac{2B\log(6/\delta)}{b\tilde{C}h^{\nu(K)}M}} = h^{\beta}\left(\frac{B}{b} + 8R\sqrt{\frac{2B\log(6/\delta)}{b\tilde{C}h^{\nu(K)+2\beta}M}} \right)$$

provided that $M \geq \frac{16B\log(6/\delta)}{b\tilde{C}h^{\nu(K)}}$. Considering (10) in the bound above gives us the claimed in the theorem bound.

It remains to check the condition $M \geq \frac{16B\log(6/\delta)}{b\tilde{C}h^{\nu(K)}}$. Since $h^{\nu} = \left(\frac{\log(6/\delta)}{M}\right)^{\frac{\nu}{2\beta+\nu}}$, this condition is equivalent to

$$M^{\frac{2\beta}{2\beta+\nu}} \geq \frac{16B}{b\tilde{C}} \big(\log(6/\delta)\big)^{\frac{2\beta}{2\beta+\nu}}.$$

Hence the condition holds whenever

$$M \geq c_0 \log(6/\delta)$$

for $c_0 > 0$ large enough. □

C Additional numerical experiments

C.1 Further plots for Section 4.1

We provide visual illustrations of the performance of the SDE estimator (11) for various choices of semi-metrics, together with the corresponding running times. In particular, the Figure 2 presents the SDE approximation accuracy reported in Table 1, and Figure 3 additionally shows the corresponding running times in seconds.

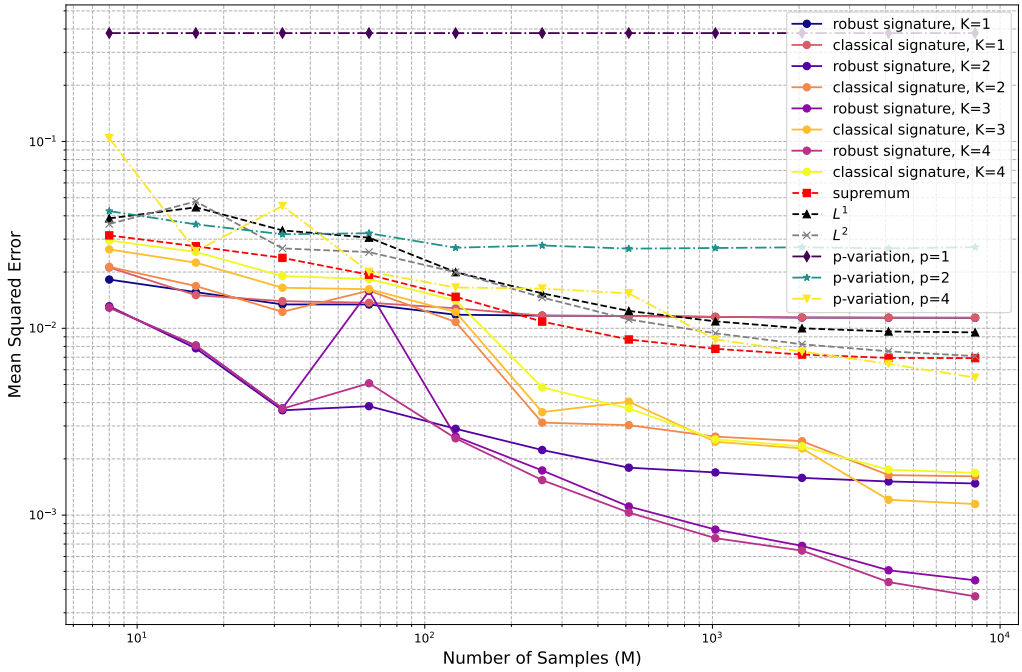

Figure 2: Regression accuracy (MSE) of the SDE estimator on the testing data as the number of training samples increases; see also Table 1.

C.2 Learning path-dependent functionals of fractional Brownian motion

In this section we present an additional numerical experiment with synthetic data. Motivated from mathematical finance, we consider the problem of learning the payoff profiles of path-dependent options. Specifically, we consider the following three path-dependent functionals

$$\Psi_i : \mathcal{X} \to \mathbb{R}, \qquad x \mapsto \begin{cases} \Psi_1(x) = \max\left(\frac{1}{T}\int_0^T x_t dt - K, 0\right) & \text{(Asian option)} \\ \Psi_2(x) = \max\left(\max_{0\le t\le T} x_t - K, 0\right) & \text{(Look-back option)} \\ \Psi_3(x) = \max(x_T - K, 0)1_{\{\max_{0\le t\le T} x_t < B\}} & \text{(Barrier option)} \end{cases}, \tag{26}$$

for some given strike, maturity and barrier $K, T, B \in \mathbb{R}_+$. In this toy example, we model the price process as a *fractional Brownian motion (fBm)* $(X_t^H)_{t\ge 0}$ with Hurst parameter $H \in (0,1)$, i.e. the unique centered, continuous Gaussian process with $X_0^H = 0$ and

$$\mathbb{E}[X_s^H X_t^H] = \frac{1}{2}(|t|^{2H} + |s|^{2H} - |t-s|^{2H}), \quad s, t \ge 0.$$

Similar to the example in Section 4.1, we can formulate this problem in the language of nonparametric regression (see Section 2) by setting $X^{(i)} = X^{H,(i)}$ and $Y_j^{(i)} = \Psi_j(X^{(i)})$, for $i = 1, \dots, M$ i.i.d. samples of the fBm and $j \in \{1, 2, 3\}$ indexing the payoff functions. We compare our nonparametric signature-based estimators both with conventional path-space norms, as before, and additionally with parametric signature-regression estimators,

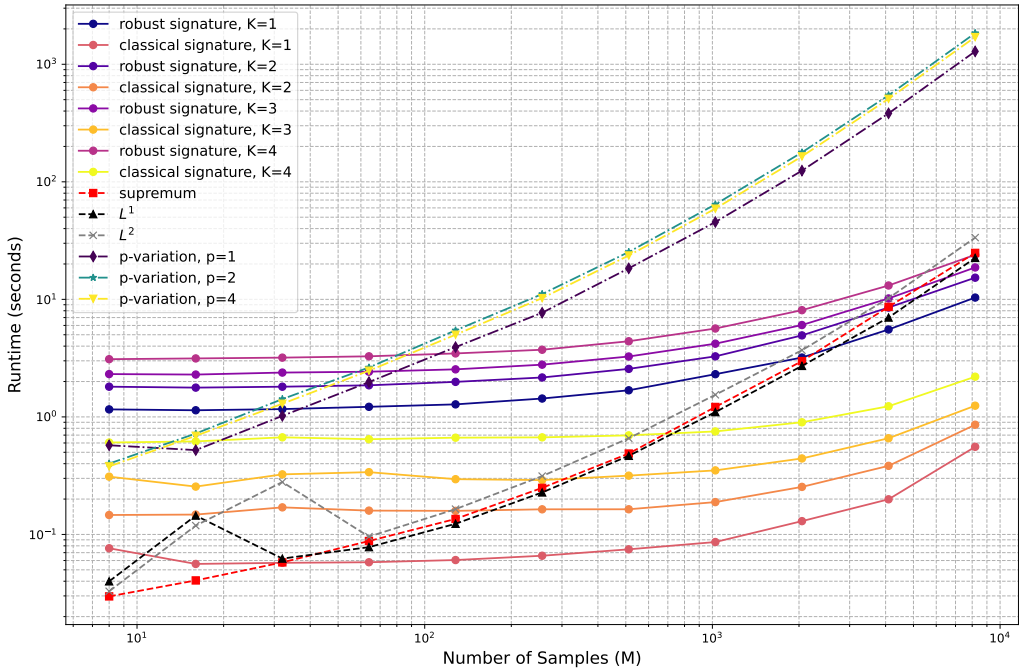

Figure 3: Running times in seconds for the experiments shown in Figure 2.

which have been employed in various applications in the literature, as mentioned in the introduction. More precisely, we compare our method with signature-based Ridge regression

$$\ell_j^\star = \operatorname*{argmin}_{\ell \in \mathcal{T}^{\leq L}} \frac{1}{M} \sum_{m=1}^{M} \left(Y_j^{(m)} - \langle \ell, \mathrm{Sig}(X^{(m)}) \rangle \right)^2 + \lambda \|\ell\|^2, \quad \lambda \in \mathbb{R}_+, \quad j \in \{1,2,3\},$$

where the Ridge penalization parameter λ is treated as a hyperparameter and L denotes the truncation level of the signature.

In the following examples, we choose the Hurst parameter to be $H = 0.3$. Similar to Table 1, in Tables 3–5 we present, for each payoff Ψ_i, an overview of the performance of the different methods across increasing sample sizes M. The experiments are conducted using the same setup as in Section 4.1, in terms of training/testing data splits and hyperparameter optimization. We exclude the p-`var` method here, as it consumes a huge amount of time compared to the other methods, as already observed in Table 1. We call the described Ridge regression method $\mathtt{R}^{\leq L}$ at the truncation level L.

M	$\mathtt{RSig}^{\leq 2}$	$\mathtt{Sig}^{\leq 2}$	$\mathtt{R}^{\leq 2}$	$\mathtt{RSig}^{\leq 3}$	$\mathtt{Sig}^{\leq 3}$	$\mathtt{R}^{\leq 3}$	$\mathtt{RSig}^{\leq 4}$	$\mathtt{Sig}^{\leq 4}$	$\mathtt{R}^{\leq 4}$	$\mathtt{RSig}^{\leq 5}$	$\mathtt{Sig}^{\leq 5}$	$\mathtt{R}^{\leq 5}$	L^1	L^2	sup
16	0.183	0.167	0.142	0.191	0.181	0.104	0.194	0.186	0.079	0.194	0.187	0.081	0.218	0.218	0.218
32	0.127	0.136	0.139	0.125	0.145	0.092	0.125	0.154	0.169	0.125	0.171	0.069	0.218	0.218	0.218
64	0.090	0.101	0.143	0.102	0.116	0.089	0.116	0.131	0.063	0.083	0.136	0.067	0.119	0.150	0.119
128	0.071	0.089	0.135	0.070	0.097	0.078	0.074	0.107	0.066	0.074	0.110	0.064	0.112	0.072	0.106
256	0.052	0.070	0.128	0.055	0.074	0.072	0.052	0.086	0.057	0.056	0.092	0.057	0.069	0.052	0.079
512	0.032	0.050	0.124	0.037	0.064	0.072	0.040	0.074	0.052	0.042	0.079	0.047	0.054	0.045	0.062
1024	0.027	0.047	0.124	0.033	0.066	0.072	0.033	0.076	0.052	0.034	0.069	0.045	0.046	0.056	0.052
2048	0.022	0.032	0.124	0.025	0.045	0.072	0.026	0.057	0.051	0.027	0.059	0.044	0.036	0.046	0.045
4096	0.015	0.030	0.123	0.019	0.038	0.072	0.020	0.049	0.051	0.021	0.053	0.045	0.029	0.038	0.034
8192	0.010	0.021	0.123	0.014	0.028	0.072	0.016	0.037	0.050	0.016	0.047	0.042	0.018	0.031	0.031
time	14.804	0.750	0.680	18.460	1.105	0.659	23.731	1.890	1.538	31.520	3.768	2.356	27.980	42.802	28.424

Table 3: Regression accuracy (RMSE) on Asian options Ψ_1 with $K = 0.5$ and $H = 0.3$. The last row corresponds to the evaluation time (in seconds) required for the whole procedure for the largest sample size $M = 8192$.

M	$\texttt{RSig}^{\leq 2}$	$\texttt{Sig}^{\leq 2}$	$\texttt{R}^{\leq 2}$	$\texttt{RSig}^{\leq 3}$	$\texttt{Sig}^{\leq 3}$	$\texttt{R}^{\leq 3}$	$\texttt{RSig}^{\leq 4}$	$\texttt{Sig}^{\leq 4}$	$\texttt{R}^{\leq 4}$	$\texttt{RSig}^{\leq 5}$	$\texttt{Sig}^{\leq 5}$	$\texttt{R}^{\leq 5}$	L^1	L^2	sup
16	0.312	0.389	0.231	0.313	0.452	0.216	0.311	0.485	0.226	0.310	0.498	0.227	0.520	0.270	0.267
32	0.285	0.326	0.249	0.370	0.388	0.238	0.333	0.420	0.217	0.315	0.436	0.222	0.239	0.231	0.245
64	0.235	0.269	0.248	0.221	0.305	0.198	0.217	0.343	0.170	0.215	0.361	0.173	0.235	0.465	0.337
128	0.228	0.252	0.235	0.217	0.276	0.192	0.212	0.298	0.165	0.210	0.309	0.160	0.233	0.218	0.304
256	0.222	0.244	0.229	0.201	0.247	0.184	0.195	0.267	0.157	0.193	0.278	0.142	0.245	0.211	0.254
512	0.212	0.222	0.227	0.188	0.233	0.182	0.181	0.241	0.153	0.179	0.251	0.143	0.215	0.204	0.223
1024	0.210	0.221	0.227	0.180	0.232	0.182	0.182	0.228	0.150	0.181	0.230	0.136	0.203	0.220	0.208
2048	0.205	0.209	0.227	0.166	0.198	0.181	0.158	0.210	0.148	0.156	0.213	0.131	0.187	0.202	0.194
4096	0.203	0.211	0.227	0.158	0.192	0.181	0.150	0.198	0.149	0.147	0.206	0.130	0.174	0.185	0.174
8192	0.201	0.205	0.227	0.154	0.181	0.181	0.144	0.183	0.148	0.141	0.196	0.130	0.164	0.172	0.168
time	15.329	0.752	0.511	18.865	1.183	0.681	24.669	2.062	1.230	32.128	3.897	2.339	25.336	49.119	32.162

Table 4: Regression accuracy (RMSE) on Look-back options Ψ_2 with $K = 0.5$ and $H = 0.3$. The last row corresponds to the evaluation time (in seconds) required for the whole procedure for the largest sample size $M = 8192$.

M	$\texttt{RSig}^{\leq 2}$	$\texttt{Sig}^{\leq 2}$	$\texttt{R}^{\leq 2}$	$\texttt{RSig}^{\leq 3}$	$\texttt{Sig}^{\leq 3}$	$\texttt{R}^{\leq 3}$	$\texttt{RSig}^{\leq 4}$	$\texttt{Sig}^{\leq 4}$	$\texttt{R}^{\leq 4}$	$\texttt{RSig}^{\leq 5}$	$\texttt{Sig}^{\leq 5}$	$\texttt{R}^{\leq 5}$	L^1	L^2	sup
16	0.092	0.092	0.093	0.093	0.093	0.144	0.090	0.090	0.547	0.088	0.091	0.393	0.099	0.099	0.099
32	0.096	0.097	0.094	0.097	0.097	0.123	0.091	0.091	0.139	0.091	0.091	0.142	0.099	0.099	0.099
64	0.077	0.098	0.093	0.094	0.097	0.090	0.098	0.098	0.090	0.098	0.098	0.090	0.106	0.109	0.093
128	0.072	0.072	0.094	0.073	0.071	0.091	0.073	0.071	0.091	0.073	0.071	0.091	0.092	0.092	0.093
256	0.074	0.073	0.094	0.074	0.074	0.090	0.074	0.075	0.091	0.075	0.076	0.091	0.093	0.092	0.093
512	0.073	0.073	0.092	0.073	0.073	0.089	0.074	0.074	0.089	0.075	0.075	0.089	0.092	0.092	0.093
1024	0.073	0.073	0.092	0.073	0.073	0.088	0.075	0.075	0.088	0.076	0.076	0.084	0.091	0.091	0.092
2048	0.072	0.072	0.092	0.069	0.069	0.088	0.069	0.069	0.087	0.068	0.069	0.084	0.092	0.090	0.092
4096	0.071	0.071	0.092	0.066	0.067	0.088	0.066	0.073	0.088	0.067	0.068	0.084	0.089	0.090	0.093
8192	0.071	0.071	0.092	0.065	0.066	0.088	0.065	0.066	0.087	0.064	0.067	0.083	0.086	0.090	0.089
time	15.214	0.769	0.512	18.598	1.257	1.054	23.668	2.043	1.239	30.942	3.919	2.313	22.830	33.232	24.010

Table 5: Regression accuracy (RMSE) on Barrier options Ψ_3 with $K = 0.5$, $B = 1.4$ and $H = 0.3$. The last row corresponds to the evaluation time (in seconds) required for the whole procedure for the largest sample size $M = 8192$.

C.3 Regression on real-world data

To complement the classification results in Section 4.2, we evaluate our methods on real-world regression datasets from the AEON time series regression benchmark. For our methods (as well as for methods with sup and L^2 metrics), we fix the signature truncation level to 3 across all datasets and select the bandwidth via 3-fold cross-validation over a grid of 50 equispaced values in [0.001,10]. Results are reported in Table 6.

On `Covid3Month`, our `Sig` estimator achieves the lowest RMSE. On `AppliancesEnergy` and `BeijingPM10Quality`, the parametric `SigP` baseline performs best, while `RSig` remains competitive. We remark once again that our method is significantly faster than the alternatives

Dataset	`DTW`	`DTWA`	`SigP`	`Sup`	L^2	`RSig`	`Sig`
Covid3Month	0.06	0.05	0.05	0.05	0.05	0.05	**0.04**
AppliancesEnergy	5.23	5.60	**2.35**	3.72	3.56	3.08	3.44
BeijingPM10Quality	141.61	141.64	**106.02**	120.55	196.29	128.03	141.40

Table 6: Test RMSE comparison across benchmark datasets. Best result in each row is shown in bold.