Random Fourier Signature Features

Csaba Tóth*, Harald Oberhauser†, and Zoltán Szabó†

Abstract. Tensor algebras give rise to one of the most powerful measures of similarity for sequences of arbitrary length called the signature kernel accompanied with attractive theoretical guarantees from stochastic analysis. Previous algorithms to compute the signature kernel scale quadratically in terms of the length and number of the sequences. To mitigate this severe computational bottleneck, we develop a random Fourier feature-based acceleration of the signature kernel acting on the inherently non-Euclidean domain of sequences. We show uniform approximation guarantees for the proposed unbiased estimator of the signature kernel, while keeping its computation linear in the sequence length and number. In addition, combined with recent advances on tensor projections, we derive two even more scalable time series features with favourable concentration properties and computational complexity both in time and memory. Our empirical results show that the reduction in computational cost comes at a negligible price in terms of accuracy on moderate size datasets, and it enables one to scale to large datasets up to a million time series. We release the code publicly available at https://github.com/tgcsaba/ksig.

Key words. Signature kernel, tensor random projections, concentration of measure, sequential data.

MSC codes. 60L10, 65C20, 68T10

1. Introduction.

Machine learning has successfully been applied to tasks that require learning from complex and structured data types on non-Euclidean domains. Feature engineering on such domains is often tackled by exploiting the geometric structure and symmetries existing within the data [5]. Learning from sequential data (such as video, text, audio, time series, health data, etc.) is a classic, but an ongoing challenge due to the following properties:

- *Non-Euclidean data.* The data domain is nonlinear since there is no obvious and natural way of adding sequences of different length.
- *Time-space patterns.* Statistically significant patterns can be distributed over time and space, that is, capturing the order structure in which "events" arise is crucial.
- *Time-warping invariance.* The meaning of many sequences is often invariant to reparametrization also frequently called time-warping, at least to an extent; e.g. a sentence spoken quicker or slower contains (essentially) the same information.
- *Discretization and irregular sampling.* Sequences often arise by sampling along an irregularly spaced grid of an underlying continuous time process. A general methodology should be robust as the sampling gets finer, sequences approximate paths (continuous-time limit), or as the discretization grid varies between sequences.
- *Scalability.* Sequence datasets can quickly become massive, so the computational complexity should grow subquadratically, in terms of all of the state-space dimension, and the length and number of sequences.

The signature kernel $\mathsf{k}_{\mathsf{Sig}}$ is the state-of-the-art kernel for sequential data [79, 65, 44] that addresses the first 4 of the above questions and can rely on the modular and powerful framework of kernel learning [66]. Its construction is motivated by classic ideas from stochastic analysis

*Mathematical Institute, University of Oxford (toth@maths.ox.ac.uk, oberhauser@maths.ox.ac.uk).
†Department of Statistics, London School of Economics (z.szabo@lse.ac.uk).

that give a structured description of a sequence by developing it into a series of tensors. We refer to [43] for a recent overview of its various constructions and applications. In the real-world, various phenomena are well-modelled by systems of differential equations. The path signature arises naturally in the context of controlled differential equations. The role of the signature here is to provide a basis for the effects of a driving signal on systems of controlled differential equations. In essence, it captures the interactions of a controlling signal with a nonlinear system. This explains the widespread applicability of signatures to various problems across the sciences [50]. There is also geometric intuition behind signatures, see Section 1.2.4 in [10].

Features vs Kernel/Primal vs Dual. Kernel learning circumvents the costly evaluation of a high- or infinite-dimensional feature map by replacing it with the computation of a Gram matrix which contains as entries the inner products of features between all pairs of data points. This can be very powerful since the inner product evaluation can often be done cheaply by the celebrated "kernel trick", even for infinite-dimensional feature spaces, but the price is that now the computational cost is quadratic in the number of samples, and downstream algorithms further often incur a cubic cost usually in the form of a matrix inversion. On the other hand, when finite-dimensional features can used for learning, the primal formulation of a learning algorithm can perform training and inference in a cost that is linear with respect to the sample size assuming that the feature dimension is fixed. This motivates the investigation of finite-dimensional approximations to kernels that mimic their expressiveness at a lower computational cost. It is an interesting question how the feature dimension should scale with the dataset size to maintain a given (optimal) learning performance in downstream tasks, which is investigated for instance by [63, 7, 72, 45, 70, 40].

Computational Cost of the Signature kernel. In the context of the signature kernel, one data point is itself a whole sequence. Hence, given a data set $\mathbf{X}$ consisting of $N \in \mathbb{Z}_+$ sequences where each sequence $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_{\ell_\mathbf{x}})$ is of maximal length $\ell_\mathbf{x} \leq \ell \in \mathbb{Z}_+$ and has sequence entries $\mathbf{x}_i$ in a state-space of dimension d, then the existing algorithms to evaluate the Gram matrix of the $\mathsf{k}_{\mathsf{Sig}}$ scale quadratically, i.e. as $O(N^2\ell^2 d)$, both in sequence length ℓ and number of sequences N. So far this has only been addressed by subsampling (either directly the sequence elements to reduce the length or by column subsampling via the Nyström approach [85]), which can lead to crude approximations and performance degradation on large-scale datasets.

Contribution. Random Fourier Features (RFF) [56] is a classic technique to enjoy both the benefits of the primal and dual approach. Here, a low-dimensional and random feature map is constructed, which although does not approximate the feature map of a translation-invariant kernel, its inner product is with high probability close to the kernel itself. The main contribution of this article is to carry out such a construction for the signature kernel. Concretely, we construct a random feature map on the domain of sequences called Random Fourier Signature Features (RFSF), such that its inner product is a random kernel $\tilde{\mathsf{k}}_{\mathsf{Sig}}$ for sequences that is both (i) an unbiased estimator for $\mathsf{k}_{\mathsf{Sig}}$, and (ii) has analogous probabilistic approximation guarantees to the classic RFF kernel. The challenge is that a direct application of the classic RFF technique is not feasible since this relies on Bochner's theorem which does not apply since the sequence domain is not even a linear space and the feature domain is non-Abelian, which makes the use of (generalizations of [27]) Bochner's theorem difficult due to the lack of sufficiently explicit representations. We tackle this challenge by combining the algebraic structure of signatures with probabilistic concentration arguments; a careful analysis of the error propagation yields

uniform concentration guarantees similar to the RFF on $\mathbb{R}^d$. Then, we introduce dimensionality reduction techniques for random tensors further approximating $\tilde{\mathsf{k}}_{\mathsf{Sig}}$ to define the extremely scalable variants $\tilde{\mathsf{k}}^{\mathsf{DP}}_{\mathsf{Sig}}$ and $\tilde{\mathsf{k}}^{\mathsf{TRP}}_{\mathsf{Sig}}$ called RFSF-DP and RFSF-TRP saving considerable amounts of computation time and memory by low-dimensional projection of the feature set of the RFSF.

Hence, analogously to the classic RFF construction, the random kernels $\tilde{\mathsf{k}}_{\mathsf{Sig}}, \tilde{\mathsf{k}}^{\mathsf{DP}}_{\mathsf{Sig}}, \tilde{\mathsf{k}}^{\mathsf{TRP}}_{\mathsf{Sig}}$ simultaneously enjoy the expressivity of an infinite-dimensional feature space as well as linear complexity in sequence length. This overcomes the arguably biggest drawback of the signature kernel, which is the quadratic complexity in sample size and sequence length; the price for reducing the complexities by an order is that this approximation only holds with high probability. As in the case of the classic RFF, our experiments show that this is in general a very attractive tradeoff. Concretely, we demonstrate in the experiments that the proposed random features (1) provide comparable performance on moderate sized datasets to full-rank (quadratic time) signature kernels, (2) outperform other random feature approaches for time series on both moderate- and large-scale datasets, (3) allow scaling to datasets of a million time series.

Related Work. The signature kernel has found many applications; for example, it is used in ABC-Bayes [21], economic scenario validation [1], amortised likelihood estimation [22], the analysis of RNNs [24], analysis of trajectories in Lie groups [42], metrics for generative modelling [6, 34], or dynamic analysis of topological structures [28]. For a general overview see [43]. All of these applications can benefit from a faster computation of the signature kernel with theoretical guarantees. Previous approaches address the quadratic complexity of the signature kernel only by subsampling in one form or another: [36] combine a structured Nyström type-low rank approximation to reduce complexity in dimension of samples and sequence length, [79] combine this with inducing point and variational methods, [65] uses sequence-subsampling, [44] use diagonal approximations to Gram matrices in a variational setting. Related to this work is also the random nonlinear projections in [49]; further, [51] combine linear dimension projection in a general pipeline and [16] use signatures in reservoir computing. Directly relevant for this work is recent progress on tensorized random projections [73, 59]. Random Fourier Features [56, 58] are well-understood theoretically [74, 69, 70, 46, 2, 75, 9, 80, 9]. In particular, its generalization properties are studied in e.g. [3, 45, 72, 40], where it is shown that the feature dimension need only scale sublinearly in the dataset size for supervised learning, and a similar result also holds for kernel principal component analysis [70]. Several variations have been proposed over the years [41, 23, 14, 88, 13, 12, 15], even finding applications in deep learning [76]. Alternative random feature approaches for polynomial and Gaussian kernels based on tensor sketching have been proposed in e.g. [84, 83, 82]. Gaussian sketching has also been applied in the RKHS for kernel approximation [38]. For a survey, the reader is referred to [9, 47].

Outline. Section 2 provides background on the prerequisites of our work: Random Fourier Features, and Signature Features/Kernels. Section 3 contains our proposed methods with theoretical results; it introduces Random Fourier Signature Features (RFSF) $\tilde{\varphi}_{\mathsf{Sig}_{\leq M}}$, RFSF kernels $\tilde{\mathsf{k}}_{\mathsf{Sig}_{\leq M}}$ (where $M \in \mathbb{Z}_+$ is the truncation level introduced later), and most importantly their theoretical guarantees. Theorem 3.2 quantifies the approximation $\mathsf{k}_{\mathsf{Sig}_{\leq M}}(\mathbf{x}, \mathbf{y}) \approx \tilde{\mathsf{k}}_{\mathsf{Sig}_{\leq M}}(\mathbf{x}, \mathbf{y})$ uniformly. Then, we discuss additional variants: the RFSF-DP kernel $\tilde{\mathsf{k}}^{\mathsf{DP}}_{\mathsf{Sig}_{\leq M}}$ and the RFSF-TRP kernel $\tilde{\mathsf{k}}^{\mathsf{TRP}}_{\mathsf{Sig}_{\leq M}}$, which build on the previous construction using dimensionality reduction with cor-

responding concentration results in Theorems 3.5 and 3.8. Section 4 compares the performance of the proposed scalable signature kernels against popular approaches on SVM multivariate time series classification, which demonstrates that the proposed kernel not only significantly improves the computational complexity of the signature kernel, it also provides comparable performance, and in some cases even improvements in accuracy as well. Hence, we take the best of both worlds: linear batch, sequence, and state-space dimension complexities, while approximately enjoying the expressivity of an infinite-dimensional RKHS with high probability.

2. Prerequisites.

Notation. We denote the real numbers by $\mathbb{R}$, natural numbers by $\mathbb{N} \coloneqq \{0, 1, 2, \dots\}$, positive integers by $\mathbb{Z}_+ \coloneqq \{1, 2, 3, \dots\}$, the range of positive integers from 1 to $n \in \mathbb{Z}_+$ by $[n] \coloneqq \{1, 2, \dots, n\}$. Given $a, b \in \mathbb{R}$, we denote their maximum by $a \vee b \coloneqq \max(a, b)$ and their minimum by $a \wedge b \coloneqq \min(a, b)$. We define the collection of all ordered m-tuples with non-repeating entries starting from 1 up to n including the endpoints by

$$\Delta_m(n) \coloneqq \{1 \leq i_1 < i_2 < \cdots < i_m \leq n \,:\, i_1, i_2, \dots, i_n \in [n]\}\,. \tag{2.1}$$

In general, $\mathcal{X}$ refers to a subset of the input domain, where the various objects are defined, generally taken to be a subset $\mathbb{R}^d$ unless otherwise stated. For a vector $\mathbf{x} \in \mathbb{R}^d$, we denote its ℓ_p norm by $\|\mathbf{x}\|_p \coloneqq \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$. For a matrix $\mathbf{A} \in \mathbb{R}^{d\times e}$, we denote the spectral and the Frobenius norm by $\|\mathbf{A}\|_2 \coloneqq \sup_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2$ and $\|\mathbf{A}\|_F \coloneqq \left(\sum_{i=1}^e \|\mathbf{A}\mathbf{e}_i\|_2^2\right)^{1/2}$, where $\{\mathbf{e}_1, \dots, \mathbf{e}_e\}$ is the canonical basis of $\mathbb{R}^e$. The transpose of a matrix $\mathbf{A}$ is denoted by $\mathbf{A}^\top$. For a differentiable $f : \mathbb{R}^d \to \mathbb{R}$, we denote its gradient at $\mathbf{x} \in \mathbb{R}^d$ by $\nabla f(\mathbf{x}) \coloneqq (\partial f(\mathbf{x})/\partial x_i)_{i=1}^d$, and its collection of partial derivatives with respect to $\mathbf{s} \coloneqq (x_{i_1}, \dots, x_{i_k})$ by $\partial_{\mathbf{s}} f(\mathbf{x}) \coloneqq \left(\partial f(\mathbf{x})/\partial x_{i_j}\right)_{j=1}^k$.

$\mathcal{X}_{\mathsf{seq}}$ refers to sequences of finite, but unbounded length with values in the set $\mathcal{X}$:

$$\mathcal{X}_{\mathsf{seq}} \coloneqq \{\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_L) : \mathbf{x}_i \in \mathcal{X}, L \in \mathbb{Z}_+\}.$$

We denote the length of a sequence $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_L) \in \mathcal{X}_{\mathsf{seq}}$ by $\ell_{\mathbf{x}} \coloneqq L$, and define the 1^{st}-order forward differencing operator as $\delta\mathbf{x}_i \coloneqq \mathbf{x}_{i+1} - \mathbf{x}_i$. We define the 1-variation functional of a sequence $\mathbf{x} \in \mathcal{X}_{\mathsf{seq}}$ as $\|\mathbf{x}\|_{1\text{-var}} \coloneqq \sum_{i=1}^{\ell_{\mathbf{x}}-1} \|\delta\mathbf{x}_i\|_2$ as a measure of sequence complexity.

Random Fourier Features. Kernel methods allow to implicitly use an infinite-dimensional feature map $\varphi : \mathcal{X} \to \mathcal{H}$ by evaluation of the inner product $\mathsf{k}(\mathbf{x}, \mathbf{y}) = \langle \varphi(\mathbf{x}), \varphi(\mathbf{y}) \rangle_{\mathcal{H}}$, when $\mathcal{H}$ is a Hilbert space. This inner product can often be evaluated without direct computation of $\varphi(\mathbf{x})$ and $\varphi(\mathbf{y})$ via the kernel trick. Although this makes them a powerful tool due to the resulting flexibility, the price of this flexibility is a trade-off in complexity with respect to the number of samples $N \in \mathbb{Z}_+$. Disregarding the price of evaluating the kernel $\mathsf{k}(\mathbf{x}, \mathbf{y})$ momentarily, kernel methods require the computation of a Gram matrix with $O(N^2)$ entries, that further incurs an $O(N^3)$ computational cost by most downstream algorithms, such as KRR [67], GP [61], and SVM [66]. Several techniques reduce this complexity, and the focal point of this article is the Random Fourier Feature (RFF) technique of [56, 57, 58], which can be applied to any continuous, bounded, translation-invariant kernel on $\mathbb{R}^d$.[1] Throughout, we write with some abuse of notation $\mathsf{k}(\mathbf{x} - \mathbf{y}) \equiv \mathsf{k}(\mathbf{x}, \mathbf{y})$. Next, we outline the RFF construction.

[1] A kernel is called translation-invariant if $\mathsf{k}(\mathbf{x}, \mathbf{y}) = \mathsf{k}(\mathbf{x} + \mathbf{z}, \mathbf{y} + \mathbf{z})$ for any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^d$.

A corollary of Bochner's theorem [64] is that any continuous, bounded, and translation-invariant kernel $\mathsf{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ can be represented as the Fourier transform of a non-negative finite measure Λ called the spectral measure associated to k, i.e. for $\mathbf{x}, \mathbf{y} \in \mathcal{X}$

$$\mathsf{k}(\mathbf{x} - \mathbf{y}) = \int_{\mathbb{R}^d} \exp(i\mathbf{w}^\top(\mathbf{x} - \mathbf{y}))\mathrm{d}\Lambda(\mathbf{w}).$$

We may, without loss of generality, assume that Λ is a probability measure such that $\Lambda(\mathbb{R}^d) = 0$, which amounts to working with the kernel $\mathsf{k}(\mathbf{x} - \mathbf{y})/\mathsf{k}(\mathbf{0})$. [56] proposed to draw $\tilde{d} \in \mathbb{Z}_+$ i.i.d. samples from Λ, $\mathbf{w}_1, \ldots, \mathbf{w}_{\tilde{d}} \overset{\text{i.i.d.}}{\sim} \Lambda$, to define the random feature map for $\mathbf{x} \in \mathcal{X}$ by

$$\tilde{\varphi} : \mathcal{X} \to \tilde{\mathcal{H}} := \mathbb{R}^{2\tilde{d}}, \quad \tilde{\varphi}(\mathbf{x}) := \frac{1}{\sqrt{\tilde{d}}} \left(\cos\left(\mathbf{W}^\top \mathbf{x}\right), \sin\left(\mathbf{W}^\top \mathbf{x}\right) \right), \tag{2.2}$$

where $\mathbf{W} = (\mathbf{w}_i)_{i=1}^{\tilde{d}} \in \mathbb{R}^{d \times \tilde{d}}$. Then, the corresponding random kernel is defined for $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ as

$$\tilde{\mathsf{k}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \quad \tilde{\mathsf{k}}(\mathbf{x}, \mathbf{y}) = \langle \tilde{\varphi}(\mathbf{x}), \tilde{\varphi}(\mathbf{y}) \rangle_{\tilde{\mathcal{H}}} = \frac{1}{\tilde{d}} \sum_{i=1}^{\tilde{d}} \cos\left(\mathbf{w}_i^\top (\mathbf{x} - \mathbf{y})\right) \tag{2.3}$$

to provide a probabilistic approximation to k. Indeed, it is a straightforward exercise to check that $\mathsf{k}(\mathbf{x}, \mathbf{y}) = \mathbb{E}\left[\tilde{\mathsf{k}}(\mathbf{x}, \mathbf{y})\right] \approx \tilde{\mathsf{k}}(\mathbf{x}, \mathbf{y})$. This approximation converges exponentially fast in $\tilde{d}$ and uniformly over compact subsets of $\mathbb{R}^d$ as proven in [56, Claim 1]. This bound was later tightened and extended to the derivatives of the kernel in the series of works [69, 75, 9], and we provide an adapted version under Theorem SM2.1 in the supplement.

Tensors and the tensor product. First, we provide a brief overview of tensors and tensor products of Hilbert spaces, which we will use to construct our feature space called the *free algebra over a Hilbert space*. The construction we adapt was first proposed by [52].

Let $\mathcal{H}_1, \ldots, \mathcal{H}_m$ be Hilbert spaces. To each element $(h_1, \ldots, h_m) \in \mathcal{H}_1 \times \cdots \mathcal{H}_m$, associate the multi-linear operator $h_1 \otimes \cdots \otimes h_m$ defined for each $(f_1, \ldots, f_m) \in \mathcal{H}_1 \times \cdots \times \mathcal{H}_m$ by

$$(h_1 \otimes \cdots \otimes h_m)(f_1, \ldots, f_m) := \langle h_1, f_1 \rangle_{\mathcal{H}_1} \cdots \langle h_m, f_m \rangle_{\mathcal{H}_m}.$$

Take the linear span of all such multi-linear operators to build the space

$$\mathcal{H}_1 \otimes' \cdots \otimes' \mathcal{H}_m := \operatorname{span}\left\{ h_1 \otimes \cdots \otimes h_m \,:\, h_1 \in \mathcal{H}_1, \ldots, h_m \in \mathcal{H}_m \right\},$$

and endow $\mathcal{H}_1 \otimes' \cdots \otimes' \mathcal{H}_m$ with an inner product via

$$\langle h_1 \otimes \cdots h_m, f_1 \otimes \cdots \otimes f_m \rangle_{\mathcal{H}_1 \otimes' \cdots \otimes' \mathcal{H}_m} := \langle h_1, f_1 \rangle_{\mathcal{H}_1} \cdots \langle h_m, f_m \rangle_{\mathcal{H}_m} \tag{2.4}$$

for all $h_1, f_1 \in \mathcal{H}_1, \ldots, h_m, f_m \in \mathcal{H}_m$, and extend by linearity to $\mathcal{H}_1 \otimes' \cdots \otimes' \mathcal{H}_m$. Taking the topological completion of this space under this inner product gives a Hilbert space denoted by $\mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_m$ called the tensor product of the Hilbert spaces $\mathcal{H}_1, \ldots, \mathcal{H}_m$. For more details about the tensor product from an algebraic point of view, see [39].

Free algebras. Now we introduce our feature space $\mathcal{H}_{\mathtt{Sig}}$. That is, we show how to embed a Hilbert space $\mathcal{H}$ into a bigger Hilbert space $\mathcal{H}_{\mathtt{Sig}}$ which is also an associative algebra[2] using a so-called free construction. Since the tensor product is associative, we can unambiguously take tensor powers of the vector space $\mathcal{H}$. Denoting $\mathcal{H}^{\otimes m} \coloneqq \mathcal{H} \otimes \cdots \otimes \mathcal{H}$, we define the free algebra over $\mathcal{H}$ as the set of sequences of tensors indexed by their degree $m \in \mathbb{N}$,

$$\bigoplus_{m\geq 0} \mathcal{H}^{\otimes m} = \left\{(t_0, \mathbf{t}_1, \mathbf{t}_2, \dots) \,:\, \mathbf{t}_m \in \mathcal{H}^{\otimes m} \text{ for } m \in \mathbb{N}, \exists n \in \mathbb{N} \text{ s.t. } N \geq n, \mathbf{t}_N = \mathbf{0}\right\}, \tag{2.5}$$

where $\bigoplus$ is the direct sum operation, $\otimes$ is the tensor product. For example, if $\mathcal{H} = \mathbb{R}^d$, then the degree-1 component is a d-dimensional vector, the degree-2 component is a $d \times d$ matrix, the degree-3 component is an array of shape $d \times d \times d$. The space $\bigoplus_{m\geq 0} \mathcal{H}^{\otimes m}$ is a vector space with addition and scalar multiplication defined for $\lambda \in \mathbb{R}$, $\mathbf{s}, \mathbf{t} \in \bigoplus_{m\geq 0} \mathcal{H}^{\otimes m}$ as

$$\mathbf{s} + \mathbf{t} \coloneqq (\mathbf{s}_m + \mathbf{t}_m)_{m\geq 0}\,, \quad \lambda\mathbf{s} \coloneqq (\lambda \mathbf{s}_m)_{m\geq 0}\,,$$

and $\mathcal{H}$ is a linear subspace of $\bigoplus_{m\geq 0} \mathcal{H}^{\otimes m}$ by identifying $\mathbf{v} \in \mathcal{H}$ as $(0, \mathbf{v}, 0, 0, \dots) \in \bigoplus_{m\geq 0} \mathcal{H}^{\otimes m}$. Further, $\bigoplus_{m\geq 0} \mathcal{H}^{\otimes m}$ is also an associative algebra since it is endowed with a (noncommutative[3]) product defined for tensors $\mathbf{s}, \mathbf{t} \in \bigoplus_{m\geq 0} \mathcal{H}^{\otimes m}$ as

$$\mathbf{st} = \left(\sum_{i=0}^{m} \mathbf{s}_i \otimes \mathbf{t}_{m-i}\right)_{m\geq 0} \in \bigoplus_{m\geq 0} \mathcal{H}^{\otimes m}.$$

This process of turning $\mathcal{H}$ into an algebra $\bigoplus_{m\geq 0} \mathcal{H}^{\otimes m}$ is a free construction; informally this means that (2.5) is the minimal structure that turns $\mathcal{H}$ into an algebra; for more details about free algebras, see [87, 62]. We now define for $\mathbf{s}, \mathbf{t} \in \bigoplus_{m\geq 0} \mathcal{H}^{\otimes m}$ their inner product as

$$\langle \mathbf{s}, \mathbf{t} \rangle_{\bigoplus_{m\geq 0} \mathcal{H}^{\otimes m}} = \sum_{m\geq 0} \langle \mathbf{s}_m, \mathbf{t}_m \rangle_{\mathcal{H}^{\otimes m}}\,,$$

where the inner product $\langle \mathbf{s}_m, \mathbf{t}_m \rangle_{\mathcal{H}^{\otimes m}}$ on $\mathcal{H}^{\otimes m}$ is as in (2.4). Finally, the completion of $\bigoplus_{m\geq 0} \mathcal{H}^{\otimes m}$ in this inner product gives a Hilbert space $\mathcal{H}_{\mathtt{Sig}}$, which is equivalently defined as

$$\mathcal{H}_{\mathtt{Sig}} = \{\mathbf{t} = (t_0, \mathbf{t}_1, \mathbf{t}_2, \dots) : \mathbf{t}_m \in \mathcal{H}^{\otimes m}, \langle \mathbf{t}, \mathbf{t} \rangle_{\mathcal{H}_{\mathtt{Sig}}} < \infty\}. \tag{2.6}$$

Path Signatures. A classic way to obtain a structured and hierarchical description of a path $\mathbf{x} : [0, T] \to \mathbb{R}^d$ is by computing a sequence of iterated integrals called the path signature of $\mathbf{x}$ given as tensors of increasing degrees $m \in \mathbb{N}$ such that the degree-m object is

$$S_m(\mathbf{x}) \coloneqq \int_{0<t_1<\cdots<t_m<T} \cdots \int \mathrm{d}\mathbf{x}(t_1) \otimes \cdots \otimes \mathrm{d}\mathbf{x}(t_m) = \int_{0<t_1<\cdots<t_m<T} \cdots \int \dot{\mathbf{x}}(t_1) \otimes \cdots \otimes \dot{\mathbf{x}}(t_m) \mathrm{d}t_1 \cdots \mathrm{d}t_m.$$

[2] An algebra A is a vector space A, where one can multiply elements together, i.e. $\mathbf{ab} \in A$ for $\mathbf{a}, \mathbf{b} \in A$.
[3] Noncommutative means that $\mathbf{ab} \neq \mathbf{ba}$ in general for elements $\mathbf{a}, \mathbf{b} \in V$ of the algebra.

Formally, we refer to the map that takes a path to its iterated integrals, $S : \mathtt{Paths} \to \mathcal{H}_{\mathtt{Sig}}$, $S(\mathbf{x}) \coloneqq (1, S_1(\mathbf{x}), S_2(\mathbf{x}), \ldots)$ as the path signature map. The domain of S is a space of paths that are regular enough such that the integrals are well-defined. Its feature space is given by applying the above construction of $\mathcal{H}_{\mathtt{Sig}}$ in (2.6) to $\mathcal{H} = \mathbb{R}^d$ with the Euclidean inner product.

Among the attractive properties of S is that it linearizes nonlinear functions of paths, that is for any continuous function f one can find a linear functional $\mathbf{w}$ of S such that

$$f(\mathbf{x}) \approx \langle \mathbf{w}, S(\mathbf{x}) \rangle \coloneqq \sum_{m \geq 0} \sum_{i_1, \ldots, i_m \in [d]} w_{i_1, \ldots, i_m} \int \dot{\mathbf{x}}^{i_1}(t_1) \cdots \dot{\mathbf{x}}^{i_m}(t_m) \mathrm{d}t_1 \cdots \mathrm{d}t_m, \tag{2.7}$$

where (2.7) $w_1, \ldots, w_d, w_{1,1}, \ldots, w_{d,d}, \ldots, w_{d,\ldots,d} \in \mathbb{R}$ denote the coordinates of $\mathbf{w}$, and the approximation holds uniformly on compacts [25, Theorem II.5] whenever the path $\mathbf{x}$ includes time as a coordinate[4]. The same results generalize to paths without an increasing coordinate up to reparametrization (i.e. time-warping) and backtracking, formally called "tree-like" equivalence, see [29]. Moreover, these iterated integrals can be well-defined beyond the setting of smooth paths; for example, the same results extend to Brownian motion, semimartingales, and even rougher paths. Rough path theory provides a systematic study that comes with a rich toolbox, that combines analytic and algebraic estimates, rich enough to cover the trajectories of large classes of stochastic processes; see [48, 26] for an introduction. Informally, iterated integrals of paths can be seen as a generalization of classical monomials and from this perspective, the approximation (2.7) can be regarded as the extension of classic polynomial regression to path-valued data. Thus at least informally it is not surprising, that vanilla signature features suffer from similar drawbacks as classic monomial features; for example, if classic monomials are replaced by other nonlinearities this often drastically improves the approximations; see e.g. [79, 65], where precomposing the signature with the RBF kernel increases learning performance.

Signature Features for Sequential Data. A challenge in machine learning when constructing feature maps for datasets of sequences is that the sequence length can vary from instance to instance; the space of sequences $\mathcal{X}_{\mathtt{seq}} = \left\{(\mathbf{x}_i)_{i=1}^{\ell} : \mathbf{x}_1, \ldots, \mathbf{x}_\ell \in \mathcal{X} \text{ and } \ell \in \mathbb{Z}_+\right\}$ includes sequences of various lengths, and they should all get mapped to the same feature space, while preserving the information about the elements themselves and their ordering. A concatenation property of path signatures called Chen's identity [50, Thm. 2.9] turns concatentation into multiplication provides a principled approach to construct features for sequences. Below we recall the construction of discrete-time signatures based on [77].

The key idea is to define the discrete-time signature of 1-step increments, and then glue features together by algebra multiplication to guarantee that the Chen identity holds by construction. Now assume we are given a static feature map $\varphi : \mathcal{X} \to \mathcal{H}$ into some Hilbert space $\mathcal{H}$. Our task is to construct from this feature map for elements of $\mathcal{X}$, a feature map for sequences of arbitrary length in $\mathcal{X}$. A natural first step is to apply the feature map φ elementwise to a sequence $\mathbf{x} \in \mathcal{X}_{\mathtt{seq}}$ to lift it to a sequence into the feature space $\mathcal{H}$ of φ, $\varphi(\mathbf{x}) \coloneqq (\varphi(\mathbf{x}_i))_{i=1}^{\ell_{\mathbf{x}}} \in \mathcal{H}_{\mathrm{seq}}$. The challenge is now to construct a feature map for sequences in $\mathcal{H}$. Simple aggregation of the individual features fails; e.g. summation of the individual features $\varphi(\mathbf{x}_i)$ would lose the order information, vectorization $(\varphi(\mathbf{x}_1), \ldots, \varphi(\mathbf{x}_{\ell_{\mathbf{x}}})) \in \mathcal{H}^{\ell_{\mathbf{x}}}$ would make

[4]This means that $\mathbf{x}^i(t) = t$ for some $i \in [d]$; more generally, a strictly increasing coordinate is sufficient.

sequences of different length not comparable. It turns out that multiplication is well-suited for this task in a suitable algebra.

Fortunately, there is a natural way to embed any Hilbert space $\mathcal{H}$ into a larger Hilbert space $\mathcal{H}_{\mathtt{Sig}}$ that is also a non-commutative algebra. First, we take the 1st-order differences,

$$\mathbf{x} \mapsto \delta\varphi(\mathbf{x}) \coloneqq (\varphi(\mathbf{x}_{i+1}) - \varphi(\mathbf{x}_i))_{i=1}^{\ell_{\mathbf{x}}-1} \in \mathcal{H}^{\ell_{\mathbf{x}}-1}, \quad \text{where } \mathbf{x} \in \mathcal{X}_{\mathtt{seq}} \tag{2.8}$$

since it is more natural to keep track of changes rather than absolute values. Then we identify $\mathcal{H}$ as a subset of $\mathcal{H}_{\mathtt{Sig}}$. The simplest choice given the above construction of $\mathcal{H}_{\mathtt{Sig}}$ is

$$\iota : \mathbf{h} \mapsto (1, \mathbf{h}, \mathbf{0}, \mathbf{0}, \ldots) \in \mathcal{H}_{\mathtt{Sig}} \quad \text{where } \mathbf{h} \in \mathcal{H}. \tag{2.9}$$

A direct calculation shows that composing the maps (2.8), (2.9), and multiplying the individual entries in $\mathcal{H}_{\mathtt{Sig}}$ results in a sequence summary using all non-contiguous subsequences, since in each multiplication step a sequence entry is either selected once or not at all. This gives rise to the discretized signatures $\varphi_{\mathtt{Sig}} : \mathcal{X}_{\mathtt{seq}} \to \mathcal{H}_{\mathtt{Sig}}$ for $\mathbf{x} \in \mathcal{X}_{\mathtt{seq}}$ with $\ell_{\mathbf{x}} \geq 2$:

$$\varphi_{\mathtt{Sig}}(\mathbf{x}) \coloneqq \prod_{i=1}^{\ell_{\mathbf{x}}-1} \iota(\delta\varphi(\mathbf{x}_i)) = \left(\sum_{\mathbf{i} \in \Delta_m(\ell_{\mathbf{x}}-1)} \delta\varphi(\mathbf{x}_{i_1}) \otimes \cdots \otimes \delta\varphi(\mathbf{x}_{i_m}) \right)_{m \geq 0}, \tag{2.10}$$

where $\Delta_m : \mathbb{Z}_+ \to \mathbb{Z}_+^m$ is as defined in (2.1) and $\mathbf{i} = (i_1, \ldots, i_m)$. Thus, the sequence feature is itself a sequence, however, now a sequence of tensors indexed by their degree $m \in \mathbb{N}$ in contrast to being indexed by the time index $i \in [\ell_{\mathbf{x}}]$. These sequence features are invariant to a natural transformation of time series called time-warping, but can also be made sensitive to it by including time as an extra coordinate with the mapping $\mathbf{x} = (\mathbf{x}_i)_{i=1}^{\ell_{\mathbf{x}}} \mapsto (t_i, \mathbf{x}_i)_{i=1}^{\ell_{\mathbf{x}}}$. It also possesses similar approximation properties to path signatures in (2.7), i.e. uniform approximation of functions of sequences on compact sets; see Appendices A and B in [77].

Despite the abstract derivation, the resulting feature map $\varphi_{\mathtt{Sig}}$ is—in principle—explicitly computable when $\mathcal{H} = \mathbb{R}^d$; see [35] for details. However, when the static feature map φ is high- or infinite-dimensional, this is not feasible and we discuss a kernel trick further below.

Remark 2.1. We used the map ι, as defined in (2.9), to embed $\mathcal{H}$ into $\mathcal{H}_{\mathtt{Sig}}$. Other choices are possible, for example one could use the embedding $\hat{\iota} : \mathcal{H} \to \mathcal{H}_{\mathtt{Sig}}$ for $\mathbf{h} \in \mathcal{H}$

$$\hat{\iota}(\mathbf{h}) \coloneqq \left(1, \mathbf{h}, \frac{\mathbf{h}^{\otimes 2}}{2!}, \frac{\mathbf{h}^{\otimes 3}}{3!}, \ldots\right) \in \mathcal{H}_{\mathtt{Sig}}. \tag{2.11}$$

This embedding is actually the classical choice in mathematics, but different choices of the embedding lead to, besides potential improvements in benchmarks, mildly different computational complexities and interesting algebraic questions [20, 77, 78].

Finally, it can be useful to only consider the first $M \in \mathbb{Z}_+$ tensors in the series $\varphi_{\mathtt{Sig}}(\mathbf{x})$ analogously to using the first M moments in classic polynomial regression to avoid overfitting. Hence, we define the M-truncated signature features for $M \in \mathbb{Z}_+$ as

$$\varphi_{\mathtt{Sig}_{\leq M}}(\mathbf{x}) \coloneqq \big(1, \varphi_{\mathtt{Sig}_1}(\mathbf{x}), \ldots, \varphi_{\mathtt{Sig}_M}(\mathbf{x}), \mathbf{0}, \mathbf{0}, \ldots\big) \quad \text{for } \mathbf{x} \in \mathcal{X}_{\mathtt{seq}},$$

where $\varphi_{\mathtt{Sig}_m}(\mathbf{x})$ is the projection of $\varphi_{\mathtt{Sig}}(\mathbf{x})$ onto $\mathcal{H}^{\otimes m}$. In practice, we regard $M \in \mathbb{Z}_+$, and the choice of the embedding as hyperparameters to optimize.

Signature Kernels. The signature is a powerful feature set for nonlinear regression on paths and sequences. A computational bottleneck associated with it is the dimensionality of the feature space $\mathcal{H}_{\mathtt{Sig}}$. As we are dealing with tensors, for $\mathcal{H}$ finite-dimensional $\varphi_{\mathtt{Sig}_m}(\mathbf{x})$ is a tensor of degree-m which has $(\dim \mathcal{H})^m$ coordinates that need to be computed. This can quickly become computationally expensive. For infinite-dimensional $\mathcal{H}$, e.g. when $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS), which is one of the most interesting settings due to the modelling flexibility, it is infeasible to directly compute $\varphi_{\mathtt{Sig}}$. In [36], the signature kernel was introduced, and it was shown that a kernel trick allows to compute the inner product of signature features up to a given degree $M \in \mathbb{Z}_+$ using dynamic programming, even when $\mathcal{H}$ is infinite-dimensional. Subsequently, [65] proposed a PDE-based algorithm to approximate the untruncated signature kernel, which was further extended in [8], and we refer to [43] for a recent overview of signature kernels. Here, we focus on discrete-time, and our starting point is the approach of [36] combined with the non-geometric approximation [20] resulting in the features (2.10).

Above we described a generic way to turn a static feature map $\varphi : \mathcal{X} \to \mathcal{H}$ into a feature map $\varphi_{\mathtt{Sig}_{\leq M}}(\mathbf{x})$ for sequences, see (2.10). The signature kernel is a powerful formalism that allows to transform any static kernel on $\mathcal{X}$ into a kernel for sequences that evolve in $\mathcal{X}$. Let $\mathsf{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous and bounded kernel, and from now on, let $\mathcal{H}$ denote its RKHS, and $\varphi(\mathbf{x}) \coloneqq \mathsf{k}_{\mathbf{x}} \equiv \mathsf{k}(\mathbf{x}, \cdot)$ the associated reproducing kernel lift for $\mathbf{x} \in \mathcal{X}$. We define the M-truncated (discretized) signature kernel $\mathsf{k}_{\mathtt{Sig}_{\leq M}} : \mathcal{X}_{\mathtt{seq}} \times \mathcal{X}_{\mathtt{seq}} \to \mathbb{R}$ for $M \in \mathbb{Z}_+$ as

$$\begin{aligned} \mathsf{k}_{\mathtt{Sig}_{\leq M}}(\mathbf{x}, \mathbf{y}) &\coloneqq \Big\langle \varphi_{\mathtt{Sig}_{\leq M}}(\mathbf{x}), \varphi_{\mathtt{Sig}_{\leq M}}(\mathbf{y}) \Big\rangle_{\mathcal{H}_{\mathtt{Sig}}} = \sum_{m=0}^{M} \big\langle \varphi_{\mathtt{Sig}_m}(\mathbf{x}), \varphi_{\mathtt{Sig}_m}(\mathbf{y}) \big\rangle_{\mathcal{H}^{\otimes m}} \\ &= \sum_{m=0}^{M} \mathsf{k}_{\mathtt{Sig}_m}(\mathbf{x}, \mathbf{y}) = \sum_{m=0}^{M} \sum_{\substack{\mathbf{i} \in \Delta_m(\ell_{\mathbf{x}}-1) \\ \mathbf{j} \in \Delta_m(\ell_{\mathbf{y}}-1)}} \delta^2_{i_1,j_1} \mathsf{k}(\mathbf{x}_{i_1}, \mathbf{y}_{j_1}) \cdots \delta^2_{i_m,j_m} \mathsf{k}(\mathbf{x}_{i_m}, \mathbf{y}_{j_m}), \end{aligned} \tag{2.12}$$

where we defined the level-m (discretized) signature kernel $\mathsf{k}_{\mathtt{Sig}_m} : \mathcal{X}_{\mathtt{seq}} \times \mathcal{X}_{\mathtt{seq}} \to \mathbb{R}$ for $m \in [M]$ as $\mathsf{k}_{\mathtt{Sig}_m}(\mathbf{x}, \mathbf{y}) \coloneqq \big\langle \varphi_{\mathtt{Sig}_m}(\mathbf{x}), \varphi_{\mathtt{Sig}_m}(\mathbf{y}) \big\rangle_{\mathcal{H}^{\otimes m}}$, and δ^2 denotes a 2^{nd}-order cross-differencing operator such that $\delta^2_{i,j}\mathsf{k}(\mathbf{x}_i, \mathbf{y}_j) \coloneqq \mathsf{k}(\mathbf{x}_{i+1}, \mathbf{y}_{j+1}) - \mathsf{k}(\mathbf{x}_{i+1}, \mathbf{y}_j) - \mathsf{k}(\mathbf{x}_i, \mathbf{y}_{j+1}) + \mathsf{k}(\mathbf{x}_i, \mathbf{y}_j)$ for $i \in [\ell_{\mathbf{x}} - 1]$ and $j \in [\ell_{\mathbf{y}} - 1]$. The key insight by [36] is equation (2.12), i.e. that $\mathsf{k}_{\mathtt{Sig}_{\leq M}}$ can be computed[5] without computing $\varphi_{\mathtt{Sig}_{\leq M}}$ itself by a kernel trick that only uses kernel evaluations.

The kernel hyperparameters are the choice of the static kernel k, for which there is a wide range of options, e.g. for $\mathcal{X} = \mathbb{R}^d$ the Gaussian, exponential or Matérn family of kernels; any hyperparameters that k comes with, such as the bandwidth; the truncation level $M \in \mathbb{Z}_+$; the choice of the algebra embedding, e.g. (2.9) or (2.11); and the choice of kernel normalization [11] that scales each level $\mathsf{k}_{\mathtt{Sig}_m}$ appropriately. It also comes with nice theoretical guarantees such as analytic estimates when sequences converge to paths, its maximum mean discrepancy (MMD) metrizes classic topologies for stochastic processes, and can lead to robust statistics in the classic statistical sense (B-robustness); see [11] for details.

Although (2.12) looks expensive to compute, [36] applies dynamic programming to efficiently compute $\mathsf{k}_{\mathtt{Sig}_{\leq M}}$ using a recursive algorithm; an alternative algorithm is the above mentioned

[5]The computation can be carried out exactly for finite M and approximately for $M = \infty$.

approach of approximating the (untruncated) signature kernel $\mathsf{k}_{\mathsf{Sig}}$ using PDE-discretization. Importantly, (2.12) avoids computing tensors, and only depends on the entry-wise evaluations of the static kernel $\mathsf{k}(\mathbf{x}_i, \mathbf{y}_j)$. Indeed, this leads to a computational cost of $O((M+d)\ell_{\mathbf{x}}\ell_{\mathbf{y}})$, which is feasible for sequences evolving in high-dimensional state-spaces, but only with moderate sequence length. Note that the same bottleneck applies to PDE-based approaches. In part, the aim of this article is to alleviate this quadratic cost in sequence length, while approximately enjoying the modelling capability of working within an infinite-dimensional RKHS.

3. Random Fourier Signature Features.

The goal of this section is to build random features for sequences, that enjoy the benefit of linear sequence length and low-dimensional feature complexity with theoretical guarantees that the corresponding inner product is close to the M-truncated (discretized) signature kernel $\mathsf{k}_{\mathsf{Sig}_{\leq M}}$ with high probability. We construct these random features in a two step process: firstly, we reduce the feature space from infinite to finite (but high) dimensionality through a careful construction using random Fourier features (RFFs), and in the second step we apply further dimensionality reduction to reduce the complexity to an even lower dimensional space in order to aid in scalability. Although we present this construction as conceptually distinct steps, the steps are coupled during the computation, and the features can be computed directly without going through the initial step.

From infinite to finite dimensions. In Section 2, we recalled the RFF construction, which associates to a continuous, bounded, translation-invariant kernel $\mathsf{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ on $\mathcal{X}$ a spectral measure Λ, and approximates k by drawing samples from Λ to define the random features $\tilde{\varphi} : \mathcal{X} \to \tilde{\mathcal{H}}$ (2.2), and the random kernel $\tilde{\mathsf{k}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (2.3). Afterwards, we presented a generic way to turn any such static features $\tilde{\varphi} : \mathcal{X} \to \tilde{\mathcal{H}}$ for elements of $\mathcal{X}$ into sequence features for sequences that evolve in $\mathcal{X}$ via $\varphi_{\mathsf{Sig}_{\leq M}} : \mathcal{X}_{\mathsf{seq}} \to \mathcal{H}_{\mathsf{Sig}}$. Applying this construction with the RFF as feature map on $\mathcal{X}$ would already result in a random feature map for sequences, i.e. a map from $\mathcal{X}_{\mathsf{seq}}$ into $\tilde{\mathcal{H}}_{\mathsf{Sig}}$. Taking the inner product in $\tilde{\mathcal{H}}_{\mathsf{Sig}}$ of this new random feature map for sequences would, however, only yield a biased estimator for the truncated signature kernel $\mathsf{k}_{\mathsf{Sig}_{\leq M}}$. We correct for this bias by revisiting our previous construction, and build an unbiased approximation to $\mathsf{k}_{\mathsf{Sig}_{\leq M}}$ using independent RFF copies in each tensor multiplication step. Then, we show in Theorem 3.2 that this random estimator comes with good probabilistic guarantees.

The probabilistic construction procedure is outlined in the following definition.

Definition 3.1. *Let $\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(M)} \overset{i.i.d.}{\sim} \Lambda^{\tilde{d}}$ be i.i.d. random matrices sampled from $\Lambda^{\tilde{d}}$ for* RFF *dimension $\tilde{d} \in \mathbb{Z}_+$, and define the independent* RFF *maps $\tilde{\varphi}_m : \mathcal{X} \to \tilde{\mathcal{H}}$ as in (2.2), i.e. $\tilde{\varphi}_m(\mathbf{x}) = \frac{1}{\sqrt{\tilde{d}}}\left(\cos(\mathbf{W}^{(m)\top}\mathbf{x}), \sin(\mathbf{W}^{(m)\top}\mathbf{x})\right)$ for $m \in [M]$ and $\mathbf{x} \in \mathcal{X}$. The M-truncated Random Fourier Signature Feature (*RFSF*) map $\tilde{\varphi}_{\mathsf{Sig}_{\leq M}} : \mathcal{X}_{\mathsf{seq}} \to \tilde{\mathcal{H}}_{\mathsf{Sig}}$ from sequences in $\mathcal{X}$ into the free algebra over $\tilde{\mathcal{H}}$ is defined for truncation level $M \in \mathbb{Z}_+$ and $\mathbf{x} \in \mathcal{X}_{\mathsf{seq}}$ as*

$$\tilde{\varphi}_{\mathsf{Sig}_{\leq M}}(\mathbf{x}) \coloneqq \left(\sum_{\mathbf{i} \in \Delta_m(\ell_{\mathbf{x}}-1)} \delta\tilde{\varphi}_1(\mathbf{x}_{i_1}) \otimes \cdots \otimes \delta\tilde{\varphi}_m(\mathbf{x}_{i_m}) \right)_{m=0}^{M}. \tag{3.1}$$

Further, the RFSF *kernel* $\tilde{\mathsf{k}}_{\mathsf{Sig}_{\leq M}} : \mathcal{X}_{\mathsf{seq}} \times \mathcal{X}_{\mathsf{seq}} \to \mathbb{R}$ *can be computed for* $\mathbf{x}, \mathbf{y} \in \mathcal{X}_{\mathsf{seq}}$ *as*

$$\begin{aligned}\tilde{\mathsf{k}}_{\mathsf{Sig}_{\leq M}}(\mathbf{x},\mathbf{y}) &:= \left\langle \tilde{\varphi}_{\mathsf{Sig}_{\leq M}}(\mathbf{x}), \tilde{\varphi}_{\mathsf{Sig}_{\leq M}}(\mathbf{y}) \right\rangle_{\tilde{\mathcal{H}}_{\mathsf{Sig}}} = \sum_{m=0}^{M} \left\langle \tilde{\varphi}_{\mathsf{Sig}_m}(\mathbf{x}), \tilde{\varphi}_{\mathsf{Sig}_m}(\mathbf{y}) \right\rangle_{\tilde{\mathcal{H}}^{\otimes m}} \\ &= \sum_{m=0}^{M} \tilde{\mathsf{k}}_{\mathsf{Sig}_m}(\mathbf{x},\mathbf{y}) = \sum_{m=0}^{M} \sum_{\substack{\mathbf{i}\in\Delta_m(\ell_{\mathbf{x}}-1)\\ \mathbf{j}\in\Delta_m(\ell_{\mathbf{y}}-1)}} \delta^2_{i_1,j_1}\tilde{\mathsf{k}}_1(\mathbf{x}_{i_1},\mathbf{y}_{j_1})\cdots\delta^2_{i_m,j_m}\tilde{\mathsf{k}}_m(\mathbf{x}_{i_m},\mathbf{y}_{j_m}),\end{aligned} \tag{3.2}$$

where we defined the level-m RFSF *kernel* $\tilde{\mathsf{k}}_{\mathsf{Sig}_m} : \mathcal{X}_{\mathsf{seq}} \times \mathcal{X}_{\mathsf{seq}} \to \mathbb{R}$ *for* $m \in \mathbb{N}$ *as* $\tilde{\mathsf{k}}_{\mathsf{Sig}_m}(\mathbf{x},\mathbf{y}) := \left\langle \tilde{\varphi}_{\mathsf{Sig}_m}(\mathbf{x}), \tilde{\varphi}_{\mathsf{Sig}_m}(\mathbf{y}) \right\rangle_{\tilde{\mathcal{H}}^{\otimes m}}$ *with the convention that* $\tilde{\mathsf{k}}_{\mathsf{Sig}_0} \equiv 1$*, and* $\tilde{\mathsf{k}}_1, \ldots, \tilde{\mathsf{k}}_M : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *are independent* RFF *kernels defined as in* (2.3) *with the random weights* $\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(M)} \in \mathbb{R}^{d\times\tilde{d}}$.

Since the feature map $\tilde{\varphi}_{\mathsf{Sig}_{\leq M}}$ can be directly evaluated in the feature space recursively, $\tilde{\mathsf{k}}_{\mathsf{Sig}_{\leq M}}$ has linear complexity in the sequence length. However, it requires computing high-dimensional tensors, where the degree-m component $\tilde{\varphi}_{\mathsf{Sig}_m}(\mathbf{x}) \in \tilde{\mathcal{H}}^{\otimes m}$ has $(\dim \tilde{\mathcal{H}})^m = (2\tilde{d})^m$ coordinates, making it infeasible for large $m, \tilde{d} \in \mathbb{Z}_+$. Remark 3.3 discusses the computational complexity in detail. Further, note that the kernel can be evaluated by means of a kernel trick exactly analogously to the evaluation of (2.12), but in this case there are no computational gains compared to the infinite-dimensional signature kernel $\mathsf{k}_{\mathsf{Sig}_{\leq M}}(\mathbf{x},\mathbf{y})$.

Next, we provide a theoretical analysis to show that the random kernel $\tilde{\mathsf{k}}_{\mathsf{Sig}_{\leq M}}(\mathbf{x},\mathbf{y})$ converges to the ground truth signature kernel $\mathsf{k}_{\mathsf{Sig}_{\leq M}}(\mathbf{x},\mathbf{y})$ exponentially fast and uniformly over compact state-spaces $\mathcal{X} \subseteq \mathbb{R}^d$, generalizing the result [56, Claim 2] to this non-Euclidean domain of sequences. Throughout the analysis, we need certain regularity properties of Λ in order to invoke quantitative versions of the law of large numbers, i.e. properties such as boundedness, existence of the moment-generating function, moment-boundedness, or belonging to certain Orlicz spaces of random variables. Boundedness of the spectral measure is too restrictive an assumption, since a continuous, bounded, translation-invariant kernel $\mathsf{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is characteristic if and only if the support of its spectral measure is $\mathbb{R}^d$, see [68, Prop. 8]. Hence, we instead work with the assumption that its moments are well-controllable, i.e. the tails of the distribution are not "too heavy". Specifically, we assume the Bernstein moment condition that

$$\mathbb{E}_{\mathbf{w}\sim\Lambda}\left[w_i^{2m}\right] \leq \frac{m! S^2 R^{m-2}}{2} \quad \text{for all } i \in [d] \tag{3.3}$$

for some $S, R > 0$. We show in the Supplementary Material under Lemmas SM1.11 and SM1.12, in a more general context, that this is equivalent to Λ being a sub-Gaussian probability measure; see e.g. [4, Sec 2.3] and [81, Sec. 2.5] about sub-Gaussianity. This of course includes the spectral measure of the Gaussian kernel defined for bandwidth $\sigma > 0$ and $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ $\mathsf{k}(\mathbf{x},\mathbf{y}) = \exp\left(-\|\mathbf{x}-\mathbf{y}\|_2^2/2\sigma^2\right)$, which has a Gaussian spectral distribution $\mathbf{w} \sim \mathcal{N}(0, 1/\sigma^2 \mathbf{I}_d)$, and therefore calculation gives $\mathbb{E}_{w\sim\mathcal{N}(0,1/\sigma^2)}\left[w^{2m}\right] = \frac{2^m\Gamma\left(m+\frac{1}{2}\right))}{\sigma^{2m}\sqrt{\pi}} < \frac{m!}{2}\left(\frac{2\sqrt{2}}{\sigma^2\sqrt[4]{\pi}}\right)^2\left(\frac{2}{\sigma^2}\right)^{m-2}$, since $\Gamma(m+1/2) < \Gamma(m+1) = m!$. Hence Λ satisfies condition (3.3) with S, R as given here. Now we state our approximation theorem regarding $\tilde{\mathsf{k}}_{\mathsf{Sig}_m}$, which quantifies that it is a (sub-)exponentially good estimator of $\mathsf{k}_{\mathsf{Sig}_m}$ with high probability and uniformly.

Theorem 3.2. *Let* $\mathsf{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ *be a continuous, bounded, translation-invariant kernel with spectral measure* Λ*, which satisfies* (3.3)*. Let* $\mathcal{X} \subset \mathbb{R}^d$ *be compact and convex with diameter* $|\mathcal{X}|$*,* $\mathcal{X}_\Delta \coloneqq \{\mathbf{x} - \mathbf{y} : \mathbf{x}, \mathbf{y} \in \mathcal{X}\}$*. Then, the following quantities are finite:* $\sigma_\Lambda^2 \coloneqq \mathbb{E}_{\mathbf{w}\sim\Lambda}\left[\|\mathbf{w}\|_2^2\right]$*,* $L \coloneqq \left\|\mathbb{E}_{\mathbf{w}\sim\Lambda}\left[\mathbf{w}\mathbf{w}^\top\right]\right\|_2^{1/2}$*,* $E_{i,j} \coloneqq \mathbb{E}_{\mathbf{w}\sim\Lambda}\left[|w_i w_j| \, \|\mathbf{w}\|_2\right]$ *and* $D_{i,j} \coloneqq \sup_{\mathbf{z}\in\mathcal{X}_\Delta} \left\|\nabla \left[\frac{\partial^2 \mathsf{k}(\mathbf{z})}{\partial z_i \partial z_j}\right]\right\|_2$ *for* $i, j \in [d]$*. Further, for any max. sequence* 1*-var* $V > 0$*, and signature level* $m \in \mathbb{Z}_+$*, for* $\epsilon > 0$

$$\mathbb{P}\left[\sup_{\substack{\mathbf{x},\mathbf{y}\in\mathcal{X}_{\mathsf{seq}} \\ \|\mathbf{x}\|_{1\text{-var}},\|\mathbf{y}\|_{1\text{-var}}\leq V}} \left|\mathsf{k}_{\mathsf{Sig}_m}(\mathbf{x},\mathbf{y}) - \tilde{\mathsf{k}}_{\mathsf{Sig}_m}(\mathbf{x},\mathbf{y})\right| \geq \epsilon\right] \leq$$

$$\leq m \begin{cases} \left(C_{d,\mathcal{X}}\left(\frac{\beta_{d,m,V}}{\epsilon}\right)^{\frac{d}{d+1}} + d\right) \exp\left(-\frac{\tilde{d}}{2(d+1)(S^2+R)}\left(\frac{\epsilon}{\beta_{d,m,V}}\right)^2\right) & \text{for } \epsilon < \beta_{d,m,V} \\ \left(C_{d,\mathcal{X}}\left(\frac{\beta_{d,m,V}}{\epsilon}\right)^{\frac{d}{(d+1)m}} + d\right) \exp\left(-\frac{\tilde{d}}{2(d+1)(S^2+R)}\left(\frac{\epsilon}{\beta_{d,m,v}}\right)^{\frac{1}{m}}\right) & \text{for } \epsilon \geq \beta_{d,m,V}, \end{cases} \tag{3.4}$$

where $C_{d,\mathcal{X}} \coloneqq 2^{\frac{1}{d+1}} 16 \, |\mathcal{X}|^{\frac{d}{d+1}} \sum_{i,j=1}^d (D_{i,j} + E_{i,j})^{\frac{d}{d+1}}$ *and* $\beta_{d,m,V} \coloneqq m\left(2V^2\left(L^2 \vee 1\right)\left(\sigma_\Lambda^2 \vee d\right)\right)^m$*.*

The proof is provided in the supplement under Theorem SM3.10. The result shows that the random kernel $\tilde{\mathsf{k}}_{\mathsf{Sig}_m}$ approximates the signature kernel $\mathsf{k}_{\mathsf{Sig}_m}$ uniformly over subsets of $\mathcal{X}_{\mathsf{seq}}$ of sequences $\mathbf{x} \in \mathcal{X}_{\mathsf{seq}}$ with maximal 1-variation V, $\|\mathbf{x}\|_{1\text{-var}} \leq V$, assuming that the state-space $\mathcal{X} \subset \mathbb{R}^d$ is a convex and compact domain. The error bound is analogous to the classic RFF bounds, in the sense that the tail probability decreases exponentially fast as a function of the RFF dimension $\tilde{d}$. The functional form of the bound is inherited from Theorem SM2.1, which provides an analogous result for the derivatives of RFF. This link follows from Lemma SM3.9, which connects the concentration of the RFSF kernel to the second derivatives of RFF.

The main difference from the classic case, i.e. [56, Claim 1] and Theorem SM2.1, is the appearance of $\beta_{d,m,V}$ which controls a regime change in the tail behaviour. Concretely, for $\epsilon < \beta_{d,m,V}$ (3.4) has a polynomial plus a sub-Gaussian tail, while for $\epsilon > \beta_{d,m,V}$ has a $(1/m)$-subexponential tail. This is not surprising as the inner summand in (3.2) is the m-fold tensor product of m independent RFF kernels, which makes the tail heavier exactly by an exponent of $1/m$. The constant itself, $\beta_{d,m,V}$, depends on (i) the maximal sequence 1-variation V, which measures a notion of time-warping invariant sequence complexity; (ii) the Lipschitz constant of the kernel L (see Examples SM3.2 and SM3.3); (iii) the trace of the second moment of Λ, $\sigma_\Lambda^2 = \mathbb{E}_{\mathbf{w}\sim\Lambda}\left[\|\mathbf{w}\|_2^2\right]$; (iv) the state-space dimension d; (v) and the signature level m itself.

Remark 3.3. Algorithm SM4.1 demonstrates the computation of the RFSF map $\tilde{\varphi}_{\mathsf{Sig}_{\leq M}}$ given a dataset of sequences $\mathbf{X} = (\mathbf{x}_i)_{i=1}^N \subset \mathcal{X}_{\mathsf{seq}}$. Upon inspection, we can deduce that the algorithm has a computational complexity of $O\left(N\ell(Md\tilde{d} + 1 + \tilde{d} + \ldots + \tilde{d}^M)\right)$. Importantly, it is linear in ℓ, the sequence length, although scales polynomially in the RFF sample size $\tilde{d}^M$.

Dimensionality Reduction: Diagonal Projection. Previously, we introduced a featurized approximation $\tilde{\mathsf{k}}_{\mathsf{Sig}_{\leq M}}$ to the signature kernel $\mathsf{k}_{\mathsf{Sig}_{\leq M}}$, called the RFSF kernel, which reduces the computation from the infinite-dimensional RKHS to a finite-dimensional feature space using random tensors. Although this makes the computation in the feature space viable of the RFSF

map $\tilde{\varphi}_{\mathsf{Sig}_{\leq M}}$, it is still tensor-valued, which incurs a computational cost of $O(\tilde{d}+\tilde{d}^2+\cdots+\tilde{d}^m)$ in the RFF dimension $\tilde{d} \in \mathbb{Z}_+$. Now, we take another step towards scalability and apply further dimensionality reduction. By examining the structure of these tensors, we introduce a diagonally projected variant called RFSF-DP that considerably reduces their sizes. We emphasize that the above RFSF construction is the crucial step: it approximates the inner product in an infinite-dimensional space, and now we further approximate it in an even lower dimensional space. The benefit is that one does not have to go through the computation of the initial RFSF map, but only the selected degrees of freedom have to be computed from the beginning.

As a first observation, we notice that the computation of (3.2) can be reformulated, due to (2.3) and linearity of the differencing operator, in the following way:

$$\tilde{\mathsf{k}}_{\mathsf{Sig}_m}(\mathbf{x},\mathbf{y}) = \frac{1}{\tilde{d}^m} \sum_{q_1,\ldots,q_m=1}^{\tilde{d}} \sum_{\substack{\mathbf{i}\in\Delta_m(\ell_\mathbf{x}-1)\\ \mathbf{j}\in\Delta_m(\ell_\mathbf{y}-1)}} \prod_{p=1}^{m} \delta^2_{i_p,j_p} \cos\left(\mathbf{w}_{q_p}^{(p)\top}(\mathbf{x}_{i_p}-\mathbf{y}_{j_p})\right) \tag{3.5}$$

by spelling out the definition of the RFF kernel, where $\mathbf{w}_1^{(1)},\ldots,\mathbf{w}_{\tilde{d}}^{(m)} \overset{\text{i.i.d.}}{\sim} \Lambda$, such that $\mathbf{W}^{(p)} = \left(\mathbf{w}_1^{(p)},\ldots,\mathbf{w}_{\tilde{d}}^{(p)}\right) \in \mathbb{R}^{d\times\tilde{d}}$ as defined in Def. 3.1. Now, we may observe that there is a dependency structure among the samples being averaged in (3.5), since the outer summation is over the Cartesian product $(q_1,\ldots,q_m) \in [\tilde{d}]^{\times m}$, which suggests that we might be able to drastically reduce the degrees of freedom by restricting this summation to only go over an independent set of samples. One way to do this is to restrict to multi-indices of the form $\mathfrak{I} \coloneqq \left\{(q,\ldots,q) \in [\tilde{d}]^{\times m} \,:\, q \in [\tilde{d}]\right\}$, i.e. we diagonally project the index set, motivating the name of the approach stated in the following definition.

Definition 3.4. *Let $\mathbf{w}_1^{(1)},\ldots,\mathbf{w}_{\tilde{d}}^{(M)} \overset{\text{i.i.d.}}{\sim} \Lambda$ for $\tilde{d} \in \mathbb{Z}_+$, and define $\hat{\varphi}_{m,q} : \mathcal{X} \to \hat{\mathcal{H}} \coloneqq \mathbb{R}^2$ with sample size $\hat{d}=1$ for $q \in [\tilde{d}]$ and $m \in [M]$, such that $\hat{\varphi}_{m,q}(\mathbf{x}) = \left(\cos(\mathbf{w}_q^{(m)\top}\mathbf{x}), \sin(\mathbf{w}_q^{(m)\top}\mathbf{x})\right)$ for $\mathbf{x} \in \mathcal{X}$. The M-truncated Diagonally Projected Random Fourier Signature Feature (RFSF-DP) map $\tilde{\varphi}^{\mathsf{DP}}_{\mathsf{Sig}_{\leq M}} : \mathcal{X}_{\mathsf{seq}} \to \tilde{\mathcal{H}}^{\mathsf{DP}}_{\mathsf{Sig}} \coloneqq \bigoplus_{m=0}^{M} \left(\hat{\mathcal{H}}^{\otimes m}\right)^{\tilde{d}}$ is defined for truncation $M \in \mathbb{Z}_+$ and $\mathbf{x} \in \mathcal{X}_{\mathsf{seq}}$ as*

$$\tilde{\varphi}^{\mathsf{DP}}_{\mathsf{Sig}_{\leq M}}(\mathbf{x}) \coloneqq \frac{1}{\sqrt{\tilde{d}}} \left(\left(\sum_{\mathbf{i}\in\Delta_m(\ell_\mathbf{x}-1)} \delta\hat{\varphi}_{1,q}(\mathbf{x}_{i_1}) \otimes \cdots \otimes \delta\hat{\varphi}_{m,q}(\mathbf{x}_{i_m}) \right)_{q=1}^{\tilde{d}} \right)_{m=0}^{M}.$$

Then, the RFSF-DP kernel can be directly computed for $\mathbf{x},\mathbf{y} \in \mathcal{X}_{\mathsf{seq}}$ via

$$\begin{aligned}
\tilde{\mathsf{k}}^{\mathsf{DP}}_{\mathsf{Sig}_{\leq M}}(\mathbf{x},\mathbf{y}) &\coloneqq \left\langle \tilde{\varphi}^{\mathsf{DP}}_{\mathsf{Sig}_{\leq M}}(\mathbf{x}), \tilde{\varphi}^{\mathsf{DP}}_{\mathsf{Sig}_{\leq M}}(\mathbf{y}) \right\rangle_{\tilde{\mathcal{H}}^{\mathsf{DP}}_{\mathsf{Sig}}} = \sum_{m=0}^{M} \left\langle \tilde{\varphi}^{\mathsf{DP}}_{\mathsf{Sig}_m}(\mathbf{x}), \tilde{\varphi}^{\mathsf{DP}}_{\mathsf{Sig}_m}(\mathbf{y}) \right\rangle_{\left(\hat{\mathcal{H}}^{\otimes m}\right)^{\tilde{d}}} \\
&= \sum_{m=0}^{M} \tilde{\mathsf{k}}^{\mathsf{DP}}_{\mathsf{Sig}_m}(\mathbf{x},\mathbf{y}) = \frac{1}{\tilde{d}} \sum_{m=0}^{M} \sum_{q=1}^{\tilde{d}} \sum_{\substack{\mathbf{i}\in\Delta_m(\ell_\mathbf{x}-1)\\ \mathbf{j}\in\Delta_m(\ell_\mathbf{y}-1)}} \delta^2_{i_1,j_1}\hat{\mathsf{k}}_{1,q}(\mathbf{x}_{i_1},\mathbf{y}_{j_1}) \cdots \delta^2_{i_m,j_m}\hat{\mathsf{k}}_{m,q}(\mathbf{x}_{i_m},\mathbf{y}_{j_m}),
\end{aligned} \tag{3.6}$$

where we defined the level-m RFSF-DP *kernel* $\tilde{\mathsf{k}}^{\mathsf{DP}}_{\mathsf{Sig}_m} : \mathcal{X}_{\mathsf{seq}} \times \mathcal{X}_{\mathsf{seq}} \to \mathbb{R}$ *for* $m \in \mathbb{N}$ *and* $\mathbf{x}, \mathbf{y} \in \mathcal{X}_{\mathsf{seq}}$ *as* $\tilde{\mathsf{k}}_{\mathsf{Sig}}(\mathbf{x}, \mathbf{y}) \coloneqq \left\langle \tilde{\varphi}^{\mathsf{DP}}_{\mathsf{Sig}_m}(\mathbf{x}), \tilde{\varphi}^{\mathsf{DP}}_{\mathsf{Sig}_m}(\mathbf{y}) \right\rangle_{\left(\hat{\mathcal{H}}^{\otimes m}\right)^{\tilde{d}}}$ *with the convention that* $\tilde{\mathsf{k}}^{\mathsf{DP}}_{\mathsf{Sig}_0} \equiv 1$*, and* $\hat{\mathsf{k}}_{m,q} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *are independent* RFF *kernels with sample size* $\hat{d} = 1$ *defined for* $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ *as* $\hat{\mathsf{k}}_{m,q}(\mathbf{x}, \mathbf{y}) \coloneqq \langle \hat{\varphi}_{m,q}(\mathbf{x}), \hat{\varphi}_{m,q}(\mathbf{y}) \rangle_{\hat{\mathcal{H}}}$ *with the random weights* $\mathbf{w}_q^{(m)} \in \mathbb{R}^d$ *for* $q \in [\tilde{d}], m \in [M]$.

Note that by the definition of the RFF kernels in (3.6), we may substitute that $\hat{\mathsf{k}}_{p,q}(\mathbf{x}, \mathbf{y}) = \cos({\mathbf{w}_q^{(p)}}^\top(\mathbf{x} - \mathbf{y}))$ for $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, so (3.6) is equivalently written for $\mathbf{x}, \mathbf{y} \in \mathcal{X}_{\mathsf{seq}}$ as

$$\tilde{\mathsf{k}}^{\mathsf{DP}}_{\mathsf{Sig}_m}(\mathbf{x}, \mathbf{y}) = \frac{1}{\tilde{d}} \sum_{q=1}^{\tilde{d}} \sum_{\substack{\mathbf{i} \in \Delta_m(\ell_{\mathbf{x}}-1) \\ \mathbf{j} \in \Delta_m(\ell_{\mathbf{y}}-1)}} \delta^2_{i_1,j_1} \cos({\mathbf{w}_q^{(1)}}^\top(\mathbf{x}_{i_1} - \mathbf{y}_{j_1})) \cdots \delta^2_{i_m,j_m} \cos({\mathbf{w}_q^{(m)}}^\top(\mathbf{x}_{i_m} - \mathbf{y}_{j_m})),$$

which is what we set out to do in the above paragraph; that is, restrict the outer summation onto the diagonal projection of the index set. Another way to look at Definition 3.4 is that the RFSF-DP kernel in (3.6) is constructed by defining $\tilde{d}$ independent RFSF kernels, each with internal RFF sample size $\hat{d} = 1$, and then taking their average; the concatenation of their corresponding features are then the features of the RFSF-DP map. Note that for RFF sample size 1, each RFF map has dimension 2, i.e. $\hat{\mathcal{H}} = \mathbb{R}^2$, and hence, the corresponding RFSF kernels have dimension $1+2+\cdots+2^M = (2^{M+1}-1)$, which by concatenation results in the overall dimensionality of the RFSF-DP kernel being $\dim \tilde{\mathcal{H}}^{\mathsf{TRP}}_{\mathsf{Sig}} = \tilde{d}\left(2^{M+1}-1\right)$. This relates to the computational complexity of the RFSF-DP map; for details see Remark 3.6.

Next, we state our concentration result regarding the level-m RFSF-DP kernel $\tilde{\mathsf{k}}^{\mathsf{DP}}_{\mathsf{Sig}_m}(\mathbf{x}, \mathbf{y})$.

Theorem 3.5. *Let* $\mathsf{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ *be a continuous, bounded, translation-invariant kernel with spectral measure* Λ*, which satisfies* (3.3)*. Then, for level* $m \in \mathbb{Z}_+$*,* $\mathbf{x}, \mathbf{y} \in \mathcal{X}_{\mathsf{seq}}$*, and* $\epsilon > 0$

$$\mathbb{P}\left[\left|\tilde{\mathsf{k}}^{\mathsf{DP}}_{\mathsf{Sig}_m}(\mathbf{x}, \mathbf{y}) - \mathsf{k}_{\mathsf{Sig}_m}(\mathbf{x}, \mathbf{y})\right| \geq \epsilon\right] \leq 2 \exp\left(-\frac{1}{4} \min\left\{\left(\frac{\sqrt{\tilde{d}}\epsilon}{2C_{d,m,\mathbf{x},\mathbf{y}}}\right)^2, \left(\frac{\tilde{d}\epsilon}{\sqrt{8}C_{d,m,\mathbf{x},\mathbf{y}}}\right)^{\frac{1}{m}}\right\}\right),$$

where $L \coloneqq \left\|\mathbb{E}_{\mathbf{w}\sim\Lambda}\left[\mathbf{w}\mathbf{w}^\top\right]\right\|$ *is the Lipschitz constant of* k*, and* $C_{d,m,\mathbf{x},\mathbf{y}} > 0$ *is bounded by*

$$C_{d,m,\mathbf{x},\mathbf{y}} \leq \sqrt{8}e^4(2\pi)^{1/4}e^{1/24}(4e^3 \left\|\mathbf{x}\right\|_{1\text{-var}} \left\|\mathbf{y}\right\|_{1\text{-var}}/m)^m \left((2d\max(S,R))^m + \left(L^2/\ln 2\right)^m\right).$$

The proof is provided in the supplement under Theorem SM3.11. The result shows that the RFSF-DP kernel converges for any two sequences $\mathbf{x}, \mathbf{y} \in \mathcal{X}_{\mathsf{seq}}$ with a $(1/m)$-subexponential convergence rate with respect to the sample size $\tilde{d} \in \mathbb{Z}_+$. Similarly to Theorem 3.2, the bound has a phase transition, where for small values of ϵ, it has a sub-Gaussian tail, while for larger values, it has a $(1/m)$-subexponential tail. A crucial difference from the previous bound is that now the phase transition happens at $\epsilon^\star = C_{d,m,\mathbf{x},\mathbf{y}} 2^{\frac{2m-3/2}{2m-1}} \tilde{d}^{\frac{1-m}{2m-1}}$, which depends on the sample size $\tilde{d}$. This means that for fixed value of $\epsilon > 0$, the phase transition always happens eventually

as $\tilde{d}$ gets large enough, hence the convergence rate with respect to $\tilde{d}$ is $(1/m)$-subexponential regardless of the value of ϵ. The slightly reduced rate of convergence compared to the RFSF kernel in Theorem 3.2 is to be expected, since the sample size of the RFSF-DP kernel is analogously reduced by an exponent of $(1/m)$ with respect to $\tilde{d}$ in comparison. The constant $C_{d,m,\mathbf{x},\mathbf{y}}$, similarly to (3.4), depends on (i) the 1-variation of sequences $\|\mathbf{x}\|_{1\text{-var}}, \|\mathbf{y}\|_{1\text{-var}}$ that measure the complexity of the sequences; (ii) $L > 0$, the Lipschitz constant of the kernel k (see Examples SM3.2, SM3.3); (iii) the moment bound parameters $S, R > 0$ from condition (3.3); (iv) the state-space dimension d; and (v) the signature level m.

Remark 3.6. Algorithm SM4.2 demonstrates the computation of the RFSF-DP map $\tilde{\varphi}^{\mathsf{DP}}_{\mathsf{Sig}_{\leq M}}$ given a dataset of sequences $\mathbf{X} = (\mathbf{x}_i)_{i=1}^N \subset \mathcal{X}_{\mathsf{seq}}$. Upon counting the operations, we deduce that the algorithm has a computational complexity $O\left(N\ell\tilde{d}(Md + 2^M)\right)$. Crucially, it is linear in both ℓ, the maximal sequence length, and $\tilde{d}$, the sample size of the random kernel.

Dimensionality Reduction: Tensor Random Projection. Previously, we built the RFSF-DP map by subsampling an independent set from the samples that constitute RFSF kernel. Here, we propose an alternative dimensionality reduction technique that starts again from the RFSF map, and uses random projections to project this generally high-dimensional tensor onto a lower dimension. Random projections are a classic technique in data science for reducing the data dimension, while preserving its important structural properties. They are built upon the celebrated Johnson-Lindenstrauss lemma [33], which states that a set of points in a high-dimensional space can be embedded into a space of much lower dimension, while approximately preserving their geometry. Exploiting this property, we construct a tensor random projected (TRP) variant of our random kernel called RFSF-TRP, such that the computation is coupled between the RFSF and TRP maps, similarly to a kernel trick.

Tensorized random projections [73, 59] construct random projections for tensors with concise parametrization that respects their tensorial nature. Given tensors $\mathbf{s}, \mathbf{t} \in \left(\mathbb{R}^d\right)^{\otimes m}$ for $m \in \mathbb{Z}_+$, the TRP map with CP (CANDECOMP/PARAFAC [37]) rank-1 is built via a random functional $\mathsf{Pr} : \left(\mathbb{R}^d\right)^{\otimes m} \to \mathbb{R}$ such that $\mathsf{Pr}(\mathbf{s}) = \langle \mathbf{p}_1 \otimes \cdots \otimes \mathbf{p}_m, \mathbf{s} \rangle_{\left(\mathbb{R}^d\right)^{\otimes m}}$, where $\mathbf{p}_1, \ldots, \mathbf{p}_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ are d-dimensional component vectors sampled from a standard normal distribution. Then, the inner product can be estimated as $\mathsf{Pr}(\mathbf{s})\mathsf{Pr}(\mathbf{t}) \approx \mathbb{E}\left[\mathsf{Pr}(\mathbf{s})\mathsf{Pr}(\mathbf{t})\right] = \langle \mathbf{s}, \mathbf{t} \rangle_{\left(\mathbb{R}^d\right)^{\otimes m}}$. Variance reduction is achieved by stacking $n \in \mathbb{Z}_+$ such random projections, each with i.i.d. component vectors $\mathbf{p}_1^{(1)}, \ldots, \mathbf{p}_m^{(n)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Hence, the TRP operator is defined as

$$\mathsf{TRP} : \left(\mathbb{R}^d\right)^{\otimes m} \to \mathbb{R}^n, \quad \mathsf{TRP}(\mathbf{s}) \coloneqq \frac{1}{\sqrt{n}} \left(\left\langle \mathbf{p}_1^{(i)} \otimes \cdots \otimes \mathbf{p}_m^{(i)}, \mathbf{s} \right\rangle \right)_{i=1}^n . \tag{3.7}$$

On the one hand, this allows to represent the random projection map onto $\mathbb{R}^n$ using only $O(nmd)$ parameters as opposed to the $O(nd^m)$ parameters in a densely parametrized random projection; and on the other, it allows for downstream computations to exploit the low-rank structure of the operator, as we shall do so in the definition stated below.

Definition 3.7. *Let $\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(M)} \overset{i.i.d.}{\sim} \Lambda^{\tilde{d}}$ be i.i.d. random matrices sampled from $\Lambda^{\tilde{d}}$ for* RFF *dimension $\tilde{d} \in \mathbb{Z}_+$, define the independent* RFF *maps $\tilde{\varphi}_m : \mathcal{X} \to \tilde{\mathcal{H}}$ as in (2.2), i.e. $\tilde{\varphi}_m(\mathbf{x}) = 1/\sqrt{\tilde{d}} \left(\cos(\mathbf{W}^{(m)\top}\mathbf{x}), \sin(\mathbf{W}^{(m)\top}\mathbf{x}) \right)$ for $m \in [M]$ and $\mathbf{x} \in \mathcal{X}$, and let $\mathbf{P}^{(1)}, \ldots, \mathbf{P}^{(M)} \overset{i.i.d.}{\sim}$*

$\mathcal{N}^{\tilde{d}}\left(\mathbf{0}, \mathbf{I}_{2\tilde{d}}\right)$ *be random matrices with i.i.d. standard normal entries. The* M*-truncated Tensor Random Projected Random Fourier Signature Feature (*RFSF-TRP*) map* $\tilde{\varphi}^{\mathsf{TRP}}_{\mathsf{Sig}_{\leq M}} \mathcal{X}_{\mathsf{seq}} \to \tilde{\mathcal{H}}^{\mathsf{TRP}}_{\mathsf{Sig}} = \mathbb{R}^{M\tilde{d}}$ *is defined for truncation level* $M \in \mathbb{Z}_+$ *and* $\mathbf{x} \in \mathcal{X}_{\mathsf{seq}}$ *as*

$$\begin{aligned}\tilde{\varphi}^{\mathsf{TRP}}_{\mathsf{Sig}_{\leq M}}(\mathbf{x}) &\coloneqq \frac{1}{\sqrt{\tilde{d}}}\left(\left(\sum_{\mathbf{i}\in\Delta_m(\ell_{\mathbf{x}}-1)}\left\langle \mathbf{p}_q^{(1)}, \delta\tilde{\varphi}_1(\mathbf{x}_{i_1})\right\rangle \cdots \left\langle \mathbf{p}_q^{(m)}, \delta\tilde{\varphi}_m(\mathbf{x}_{i_m})\right\rangle\right)_{q=1}^{\tilde{d}}\right)_{m=0}^{M} \\ &= \frac{1}{\sqrt{\tilde{d}}}\left(\sum_{\mathbf{i}\in\Delta_m(\ell_{\mathbf{x}}-1)}\left(\mathbf{P}^{(1)\top}\delta\tilde{\varphi}_1(\mathbf{x}_{i_1})\right)\odot\cdots\odot\left(\mathbf{P}^{(m)\top}\delta\tilde{\varphi}_m(\mathbf{x}_{i_m})\right)\right)_{m=0}^{M}, \qquad (3.8)\end{aligned}$$

where $\mathbf{P}^{(m)} = \left(\mathbf{p}_q^{(m)}\right)_{q=1}^{\tilde{d}} \in \mathbb{R}^{2\tilde{d}\times\tilde{d}}$*, and* $\odot$ *denotes the Hadamard product*[6]*. The* RFSF-TRP *kernel* $\tilde{\mathsf{k}}^{\mathsf{TRP}}_{\mathsf{Sig}_{\leq M}} : \mathcal{X}_{\mathsf{seq}} \times \mathcal{X}_{\mathsf{seq}} \to \mathbb{R}$ *can then be directly computed for sequences* $\mathbf{x}, \mathbf{y} \in \mathcal{X}_{\mathsf{seq}}$ *by*

$$\begin{aligned}\tilde{\mathsf{k}}^{\mathsf{TRP}}_{\mathsf{Sig}_{\leq M}}(\mathbf{x},\mathbf{y}) &\coloneqq \left\langle \tilde{\varphi}^{\mathsf{TRP}}_{\mathsf{Sig}_{\leq M}}(\mathbf{x}), \tilde{\varphi}^{\mathsf{TRP}}_{\mathsf{Sig}_{\leq M}}(\mathbf{y})\right\rangle_{\tilde{\mathcal{H}}_{\mathsf{Sig}}} = \sum_{m=0}^{M}\left\langle \tilde{\varphi}^{\mathsf{TRP}}_{\mathsf{Sig}_m}(\mathbf{x}), \tilde{\varphi}^{\mathsf{TRP}}_{\mathsf{Sig}_m}(\mathbf{y})\right\rangle_{\tilde{\mathcal{H}}^{\otimes m}} \\ &= \sum_{m=0}^{M}\tilde{\mathsf{k}}^{\mathsf{TRP}}_{\mathsf{Sig}_m}(\mathbf{x},\mathbf{y}) = \frac{1}{\tilde{d}}\sum_{m=0}^{M}\sum_{q=1}^{\tilde{d}}\sum_{\substack{\mathbf{i}\in\Delta_m(\ell_{\mathbf{x}}-1)\\ \mathbf{j}\in\Delta_m(\ell_{\mathbf{y}}-1)}}\prod_{p=1}^{m}\left\langle \mathbf{p}_q^{(p)}, \delta\tilde{\varphi}_p(\mathbf{x}_{i_p})\right\rangle\left\langle \mathbf{p}_q^{(p)}, \delta\tilde{\varphi}_p(\mathbf{y}_{j_p})\right\rangle, \qquad (3.9)\end{aligned}$$

where we defined the level-m RFSF-TRP *kernel* $\tilde{\mathsf{k}}^{\mathsf{TRP}}_{\mathsf{Sig}_m} : \mathcal{X}_{\mathsf{seq}} \times \mathcal{X}_{\mathsf{seq}} \to \mathbb{R}$ *for* $m \leq M$ *as* $\tilde{\mathsf{k}}^{\mathsf{TRP}}_{\mathsf{Sig}_m}(\mathbf{x},\mathbf{y}) \coloneqq \left\langle \tilde{\varphi}^{\mathsf{TRP}}_{\mathsf{Sig}_m}(\mathbf{x}), \tilde{\varphi}^{\mathsf{TRP}}_{\mathsf{Sig}_m}(\mathbf{y})\right\rangle_{\tilde{\mathcal{H}}^{\otimes m}}$ *with the convention that* $\tilde{\mathsf{k}}^{\mathsf{TRP}}_{\mathsf{Sig}_0} \equiv 1$.

We remark that (3.8) is equivalent to the TRP operator (3.7) applied to the RFSF map (3.1) by exploiting bilinearity of the inner product, and using that it factorizes over the tensor components, as described in (2.4). Then, the unbiasedness of (3.9) follows from the fact that the TRP operator is an isometry under expectation, which is applied to the RFSF tensor $\tilde{\varphi}_{\mathsf{Sig}_m}$, therefore $\tilde{\mathsf{k}}^{\mathsf{TRP}}_{\mathsf{Sig}_m}$ kernel is conditionally an unbiased estimator of $\tilde{\mathsf{k}}_{\mathsf{Sig}_m}$ given the RFSF weights $\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(m)} \in \mathbb{R}^{d\times\tilde{d}}$. By the tower rule for expectations, $\tilde{\mathsf{k}}^{\mathsf{TRP}}_{\mathsf{Sig}_m}$ is an unbiased estimator of $\mathsf{k}_{\mathsf{Sig}_m}$. The approximation quality is then governed by two factors: (i) how well the TRP projected kernel $\tilde{\mathsf{k}}^{\mathsf{TRP}}_{\mathsf{Sig}_m}$ approximates $\tilde{\mathsf{k}}_{\mathsf{Sig}_m}$; (ii) the quality of the approximation of $\tilde{\mathsf{k}}_{\mathsf{Sig}_m}$ with respect to $\mathsf{k}_{\mathsf{Sig}_m}$. Note that (ii) has already been discussed in Theorem 3.2. Here, we state the following theoretical result which quantifies (i). Combining these two results by means of triangle inequality and union bounding quantifies that $\tilde{\mathsf{k}}^{\mathsf{TRP}}_{\mathsf{Sig}_m}$ is a good estimator of $\mathsf{k}_{\mathsf{Sig}_m}$.

Theorem 3.8. *Let* $\mathsf{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ *be a continuous, bounded, translation-invariant kernel with spectral measure* Λ*, which satisfies* (3.3)*. Then, the following bound holds for* RFSF-TRP

[6]The Hadamard product stands for component-wise multiplication of the vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\mathbf{x}\odot\mathbf{y} = (x_i y_i)_{i=1}^n$.

kernel for signature level $m \in \mathbb{Z}_+$ *sequences* $\mathbf{x}, \mathbf{y} \in \mathcal{X}_{\mathsf{seq}}$ *and* $\epsilon > 0$

$$\mathbb{P}\left[\left|\tilde{\mathsf{k}}^{\mathsf{TRP}}_{\mathsf{Sig}_m}(\mathbf{x},\mathbf{y}) - \tilde{\mathsf{k}}_{\mathsf{Sig}_m}(\mathbf{x},\mathbf{y})\right| \geq \epsilon\right] \leq C_{d,\Lambda} \exp\left(-\left(\frac{m^2 \tilde{d}^{\frac{1}{2m}} \epsilon^{\frac{1}{m}}}{2\sqrt{2}e^3 R \left\|\mathbf{x}\right\|_{1\text{-var}} \left\|\mathbf{y}\right\|_{1\text{-var}}}\right)^{\frac{1}{2}}\right), \tag{3.10}$$

where the absolute constant is defined as $C_{d,\Lambda} \coloneqq 2\left(1 + \frac{S}{2R} + \frac{S^2}{4R^2}\right)^d$.

The proof is given in the supplement under Theorem SM3.12 utilizing the hypercontractivity of Gaussian polynomials [32] that is used to quantify the concentration of the TRP estimator. The concentration of the RFSF-TRP kernel is then governed by Theorems 3.2 and 3.8 combined. Together, they show that for smaller values of ϵ (i.e. the regime change as discussed below Theorem 3.2), the probability has a polynomial plus a sub-Gaussian tail, while for large ϵ, it has a $(\frac{1}{2m})$-subexponential tail due to (3.10), and the dominant convergence rate with respect to $\tilde{d}$ is $(\frac{1}{4m})$-subexponential. This means that in terms of convergence, RFSF-TRP is the slowest among the 3 variations introduced so far. However, it is also the most efficient in terms of overall dimension, hence downstream computational complexity as well, since $\tilde{\mathcal{H}}^{\mathsf{TRP}}_{\mathsf{Sig}} = \mathbb{R}^{M\tilde{d}}$. Remark 3.9 discusses the computational complexity in detail.

Remark 3.9. Algorithm SM4.3 demonstrates the computation of the RFSF-TRP map $\tilde{\varphi}^{\mathsf{TRP}}_{\mathsf{Sig}_{\leq M}}$ given a dataset of sequences $\mathbf{X} = (\mathbf{x}_i)_{i=1}^N \subset \mathcal{X}_{\mathsf{seq}}$. Counting the operations, here we can deduce that the algorithm has an $O\left(MN\ell\tilde{d}(d+\tilde{d})\right)$ computational complexity. This variation is also linear in ℓ, the maximal sequence length, although it is quadratic in $\tilde{d}$.

Numerical evaluation. Here, we numerically evaluate the approximation error of the proposed scalable kernels, that is, RFSF-DP and RFSF-TRP. We do not include RFSF since its dimensionality shows polynomial explosion in the base sample size $\tilde{d}$ due to its tensor-based representation, which makes its computation infeasible for reasonable values of $\tilde{d}$. We generate d-dimensional synthetic time series of length-ℓ using a VAR(1) process $\tilde{\mathbf{x}} \in \mathcal{X}_{\mathsf{seq}}$, such that $\tilde{\mathbf{x}}_0 = \mathbf{0}$ and $\tilde{\mathbf{x}}_{t+1} = {}^1\!/\!\sqrt{d} A\tilde{\mathbf{x}}_t + \epsilon_t$, where $A \sim \mathcal{N}^{d\times d}(0,1)$ and $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_d)$. Then, we compute the normalized version $\mathbf{x} \in \mathcal{X}_{\mathsf{seq}}$ of $\tilde{\mathbf{x}}$, which is rescaled to have 1-variation $V > 0$, i.e. $\mathbf{x}_t = V\tilde{\mathbf{x}}_t / \left\|\tilde{\mathbf{x}}\right\|_{1\text{-var}}$. We set $d = 10$, $\ell = 100$, $\sigma = 0.1$ and $V = 100$. We compute the squared deviation between the groundtruth signature kernel and the randomized approximations for two randomly sampled time series in this way. This process is repeated for 100 randomly sampled time series and 100 times resampled random kernel evaluations, giving rise to overall 10000 evaluations for each value of $\tilde{d}$. Figure 1 shows the average approximation error plotted against values of $\tilde{d}$ on a log-log plot. We can observe that both RFSF-DP and RFSF-TRP have approximately the same

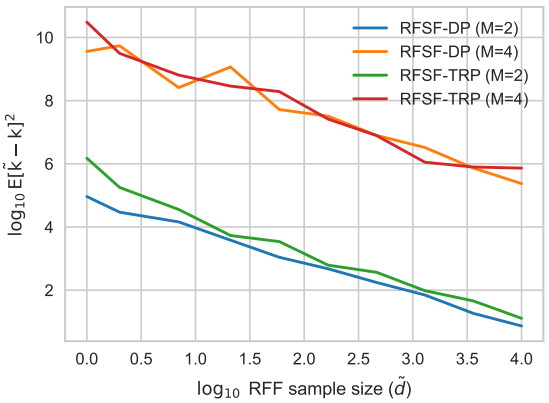

Figure 1: Approximation error of random kernels against RFF sample size on log-log plot.

Table 1: Computational complexities of kernels in our experiments; $N \in \mathbb{Z}_+$ is the number of time series, $\ell \in \mathbb{Z}_+$ is their length, $d \in \mathbb{Z}_+$ is their state-space dimension, $M \in \mathbb{Z}_+$ is the signature truncation level, $\tilde{d} \in \mathbb{Z}_+$ is the RF dimension, $W \in \mathbb{Z}_+$ is the warping length in RWS.

RFSF-DP	RFSF-TRP	KSig	KSigPDE	RWS	GAK	RBF	RFF
$O\left(N\ell\tilde{d}\left(Md+2^M\right)\right)$	$O\left(N\ell M\tilde{d}\left(d+\tilde{d}\right)\right)$	$O\left(N^2\ell^2\left(M+d\right)\right)$	$O\left(N^2\ell^2 d\right)$	$O\left(N\ell W d\right)$	$O\left(N^2\ell^2 d\right)$	$O\left(N^2\ell d\right)$	$O\left(N\ell d\tilde{d}\right)$

error curves for a given value of truncation level M, and the steepness appears to be the same across different levels of M. This means that RFSF-TRP is slightly more efficient in terms of dimensionality, since its dimension is $M\tilde{d}$ as opposed to $2^{M+1}\tilde{d}$ in RFSF-DP. We also observe that both curves are close to being linear, which indicates that the approximation error scales approximately as $O(\tilde{d}^{-\alpha})$ for some value of $\alpha > 0$.

4. Experiments.

Time series classification. We perform multivariate time series classification to investigate the performance of the scalable RFSF variants compared to the full-rank signature kernel and other quadratic time baseline kernels, and further, to demonstrate the scalability to large-scale datasets, where the quadratic sample complexity becomes prohibitive. We use support vector machine (SVM) [71] classification for classifying multivariate time series on datasets of various sizes. For quadratic time kernels, the dual SVM formulation is used, while for kernels with feature representations, we use the primal formulation that has linear complexity in the size of the dataset $n \in \mathbb{Z}_+$ aiding in scalability to truly large-scale datasets. For each considered kernel/feature, we use a GPU-based implementation provided in the KSig library[7]. For large-scale experiments with the featurized kernels, linear SVM implementation is used from the cuML library [60], while the dual SVM on moderate-scale datasets uses the sklearn library [54]. For multi-class problems, we use the one-vs-one classification strategy. This study is also the largest scale comparison of signature kernels to date which extends the datasets considered in [65]. The hardware used was 2 computer clusters equipped with overall 8 NVIDIA 3080 Ti GPUs.

Methods. We compare the proposed variants RFSF-DP and RFSF-TRP to the baselines described here: (1) the M-truncated Signature Kernel [36] KSig formulated via the kernel trick, and is a quadratic time baseline; (2) the Signature-PDE Kernel [65] KSigPDE, which uses the 2^{nd}-order PDE solver and also has quadratic complexity; (3) the Global Alignment Kernel [17] GAK, one of the most popular sequence kernels to day and can can be related to the signature kernel, see [36, Sec. 5]; (4) Random Warping Series [86] RWS, which produces features by DTW alignments between the input and randomly sampled time series; (5) the RBF kernel, which treats the whole time series as a vector of length $\mathbb{R}^{\ell d}$, (6) Random Fourier Reatures [56] RFF, which also treats the time series as a long vector. We excluded RFSF from the comparison, as it is unfeasible to compute it with reasonable sample sizes $\tilde{d}$ due to the polynomial explosion of dimensions in its tensor-based representation. The complexities are compared in Table 1.

Hyperparameter selection. For each dataset-kernel, we perform cross-validation to select the optimal hyperparameters that are optimized over the Cartesian product of the following

[7] https://github.com/tgcsaba/KSig

options. For each method that requires a static kernel, we use the RBF kernel with bandwidth hyperparameter $\sigma > 0$. This is specified in terms of a rescaled median heuristic, i.e.

$$\sigma = \alpha \operatorname{med}\left\{ \left\| \mathbf{x}_i - \mathbf{x}'_j \right\|_2 / 2 \,:\, i \in [\ell_{\mathbf{x}}], j \in [\ell_{\mathbf{x}'}], \mathbf{x}, \mathbf{x}' \in \mathbf{X} \right\}, \quad \text{for } \alpha > 0, \tag{4.1}$$

where α is chosen from $\alpha \in \{10^{-3}, \dots, 10^3\}$ on a logarithmic grid with 19 steps. For each kernel that is not normalized by default (i.e. the GAK and RBF kernels are normalized, the former is because without normalization it blows up) , we select whether to normalize to unit norm in feature space via $\mathsf{k}(\mathbf{x}, \mathbf{y}) \mapsto \mathsf{k}(\mathbf{x}, \mathbf{y}) / \sqrt{\mathsf{k}(\mathbf{x}, \mathbf{x})\mathsf{k}(\mathbf{y}, \mathbf{y})}$. The SVM hyperparameter $C > 0$ is selected from $C \in \{10^0, 10^1, 10^2, 10^3, 10^4\}$. Further, motivated by previous work that investigates the effect of path augmentations in the context of signature methods [51], we chose 3 augmentations to cross-validate over. First is parametrization encoding, which gives the classifier the ability to remove the warping invariance of a given sequence kernel, adding the time index as an additional coordinate, i.e. for each time series in the dataset $\mathbf{x} \in \mathbf{X}$, we augment it via $\mathbf{x} = (\mathbf{x}_i)_{i=1}^{\ell_{\mathbf{x}}} \mapsto (\beta i / \ell_{\mathbf{x}}, \mathbf{x}_i)_{i=1}^{\ell_{\mathbf{x}}}$, where $\beta > 0$ is the parametrization intensity chosen from $\beta \in \{10^0, 10^1, 10^2, 10^3, 10^4\}$. The second augmentation is the basepoint encoding, the role of which is to remove the translation invariance of signature features. Note that when the static base kernel is chosen to be a nonlinear kernel other than the Euclidean inner product, the signature kernel is not completely translation-invariant due to the state-space nonlinearities, but it is close being that by the L-Lipschitz property in Lemma SM3.2 valid for of the static kernels considered in this work. The basepoint encoding adds an initial $\mathbf{0}$ step at the beginning of each time series, i.e. for $\mathbf{x} \in \mathbf{X}$, $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_{\ell_{\mathbf{x}}}) \mapsto (\mathbf{0}, \mathbf{x}_1, \dots, \mathbf{x}_{\ell_{\mathbf{x}}})$. The third augmentation is the lead-lag map, which is defined as $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_{\ell_{\mathbf{x}}}) \mapsto ((\mathbf{x}_1, \mathbf{x}_1), (\mathbf{x}_2, \mathbf{x}_1), (\mathbf{x}_2, \mathbf{x}_2), \dots, (\mathbf{x}_{\ell_{\mathbf{x}}}, \mathbf{x}_{\ell_{\mathbf{x}}-1}), (\mathbf{x}_{\ell_{\mathbf{x}}}, \mathbf{x}_{\ell_{\mathbf{x}}}))$. For the truncation-based signature kernels, we select the truncation level $M \in \mathbb{Z}_+$ from $M \in \{2, 3, 4, 5\}$. For RWS, we select the warping length from $W \in \{10, 20, \dots, 100\}$ as suggested by the authors. This makes RWS the most expensive feature-based kernel, and so as to fit within the same resource limitations, we omit cross-validating over the path augmentations. We select the standard deviation $\sigma > 0$ of the warping series from the same grid as α in (4.1). For all RF approaches, we set the RF dimension $\tilde{d} \in \mathbb{Z}_+$, so the overall dimension is 1000. Note that for RFSF-DP and RFSF-TRP this is respectively $2^{M+1}\tilde{d}$ and $M\tilde{d}$, where $\tilde{d}$ is the base RFF sample size; for RWS it is the number of warping series $\tilde{d}$; while for RFF it is twice the number of samples $2\tilde{d}$.

Datasets: UEA Archive. The UEA archive [18] is a collection of overall 30 datasets for benchmarking classifiers on multivariate time series classification problems containing both binary and multi-class tasks. The data modality ranges from various sources e.g. human activity recognition, motion classification, ECG classification, EEG/MEG classification, audio spectra recognition, and others. The sizes of the datasets in terms of number of time series range from moderate (≤ 1000 examples) to large (≤ 30000), and includes various lengths between 8 and 18000. A summary of the dataset characteristics can be found in Table 2 in [18]. Pre-specified train-test splits are provided for each dataset, which we follow. We evaluate all considered kernels on the moderate datasets (≤ 1000 time series), but because the non-feature-based become very expensive computationally beyond these sizes, we only evaluate feature-based approaches on medium and large datasets (≥ 1000 time series). Each featurized approach is trained and evaluated 5 times on each dataset in order to account for the randomness in the hyperparameter selection procedure and evaluation.

Table 2: Comparison of SVM test accuracies on moderate multivariate time series classification datasets. For each row, the best result is highlighted in **bold**, and the second best in *italic*.

	RFSF-DP	RFSF-TRP	KSig	KSigPDE	RWS	GAK	RBF	RFF
ArticularyWordRecognition	0.984	0.981	**0.990**	0.983	*0.987*	0.977	0.977	0.978
AtrialFibrillation	0.373	0.320	*0.400*	0.333	**0.427**	0.333	0.267	0.373
BasicMotions	**1.000**	**1.000**	**1.000**	**1.000**	*0.995*	**1.000**	0.975	0.860
Cricket	0.964	0.964	0.958	*0.972*	**0.978**	0.944	0.917	0.886
DuckDuckGeese	0.636	*0.664*	**0.700**	0.480	0.492	0.500	0.420	0.372
ERing	0.921	0.936	0.841	*0.941*	**0.945**	0.926	0.937	0.915
EigenWorms	*0.817*	**0.837**	0.809	0.794	0.623	0.511	0.496	0.443
Epilepsy	**0.949**	*0.942*	**0.949**	0.891	0.925	0.870	0.891	0.777
EthanolConcentration	0.457	0.439	**0.479**	*0.460*	0.284	0.361	0.346	0.325
FingerMovements	0.608	0.624	**0.640**	*0.630*	0.612	0.500	0.620	0.570
HandMovementDirection	*0.573*	0.568	**0.595**	0.527	0.403	**0.595**	0.541	0.454
Handwriting	0.434	0.400	0.479	0.409	**0.591**	*0.481*	0.307	0.249
Heartbeat	0.717	0.712	0.712	**0.722**	0.714	0.717	0.717	*0.721*
JapaneseVowels	0.978	0.978	**0.986**	**0.986**	0.955	*0.981*	*0.981*	0.979
Libras	0.898	**0.928**	*0.922*	0.894	0.837	0.767	0.800	0.800
MotorImagery	*0.516*	**0.526**	0.500	0.500	0.508	0.470	0.500	0.482
NATOPS	0.906	0.908	0.922	**0.928**	*0.924*	0.922	0.917	0.900
PEMS-SF	0.800	0.808	0.827	*0.838*	0.701	**0.855**	**0.855**	0.770
RacketSports	0.874	0.861	**0.921**	*0.908*	0.878	0.849	0.809	0.755
SelfRegulationSCP1	0.868	0.856	*0.904*	*0.904*	0.829	**0.915**	0.898	0.885
SelfRegulationSCP2	0.489	0.510	*0.539*	**0.544**	0.481	0.511	0.439	0.492
StandWalkJump	0.387	0.333	*0.400*	*0.400*	0.347	0.267	**0.533**	0.267
UWaveGestureLibrary	0.882	0.881	**0.912**	0.866	*0.897*	0.887	0.766	0.846
Avg. acc.	*0.740*	0.738	**0.756**	0.735	0.710	0.702	0.692	0.656
Avg. rank	3.609	3.739	**2.348**	*2.957*	3.957	4.174	4.913	5.913

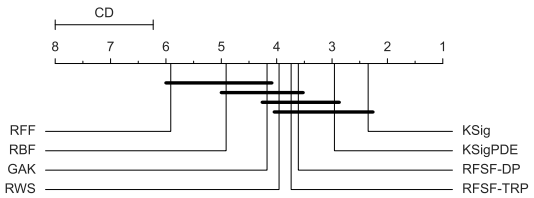

Figure 2: Critical difference diagram comparison on moderate datasets of considered approaches using two-tailed Nemenyi test [19].

Table 3: Comparison of accuracies on large-scale datasets of random features.

	RFSF-DP	RFSF-TRP	RWS	RFF
CharacterTrajectories	*0.990*	*0.990*	**0.991**	0.989
FaceDetection	*0.653*	**0.656**	0.642	0.572
InsectWingbeat	*0.436*	**0.459**	0.227	0.341
LSST	0.589	*0.624*	**0.631**	0.423
PenDigits	*0.983*	0.982	**0.989**	0.980
PhonemeSpectra	*0.204*	*0.204*	**0.205**	0.083
SITS1M	**0.745**	*0.740*	0.610	0.718
SpokenArabicDigits	**0.981**	*0.980*	**0.981**	0.964
fNIRS2MW	**0.659**	*0.658*	0.621	0.642
Avg. acc.	*0.693*	**0.699**	0.655	0.635
Avg. rank	**1.778**	*1.889*	2.222	3.333

Datasets: Mental Workload Intensity Classification. We evaluate featurized approaches on a large-scale brain-activity recording data set called `fNIRS2MW`.[8] This dataset contains brain activity recordings collected from overall 68 participants during a 30-60 minute experimental session, where they were asked to carry out tasks of varying intensity. The collected time series are sliced into 30 second segments using a sliding window, and each segment is labelled with an intensity level (0-3), giving rise to overall ~ 100000 segments, which we split in a ratio of $80-20$ for training and testing. We convert the task into a binary classification problem by assigning a label whether the task is low (0 or 1) or high (2 or 3) intensity.

Datasets: Satellite Image Classification. As a massive scale task, we use a satellite imagery dataset[9] of $N = 10^6$ time series. Each length $\ell = 46$ time series corresponds to a vegetation index calculated from remote sensing data, and the task is to classify land cover types [55] by mapping vegetation profiles to various types of crops and forested areas corresponding to 24 classes. We split the dataset in a ratio of 90-10 for training and testing.

4.1. Results. Table 2 compares test accuracies on moderate size multivariate time series classification datasets with $N \leq 1000$ from the UEA archive. `KSig` provides state-of-the-art performance among all sequence kernels with taking the highest aggregate score in terms of all of average accuracy, average rank, and number of first places. Our proposed random feature variants `RFSF-DP` and `RFSF-TRP` provide comparable performance on most of the datasets in terms of accuracy, and they are only outperformed by `KSig` and `KSigPDE` with respect to average accuracy and rank. Interestingly, `RFSF-TRP` has more first place rankings, but `RFSF-DP` performs slightly better on average. This shows that on datasets of these sizes, using either of `RFSF-DP` and `RFSF-TRP` does not sacrifice model performance - even leading to improvements in some cases, potentially due to the implicit regularization effect of restricting to a finite-dimensional feature space - and it can already provide speedups. We visualize the critical difference diagram comparison of all considered approaches in Figure 2.

Table 3 demonstrates the performance of scalable approaches, i.e. `RFSF-DP`, `RFSF-TRP`, `RWS` and `RFF` on the remaining UEA datasets ($N \geq 1000$), the dataset `fNIRS2MW` ($N = 10^5$), and the satellite dataset `SITS1M` ($N = 10^6$). We find it infeasible to perform full cross-validation for quadratic time kernels on these datasets due to expensive kernel computations and downstream cost of dual `SVM`. The results show that both variants `RFSF-DP` and `RFSF-TRP` perform significantly better on average with respect to accuracy and rank then both `RWS` and `RFF`. Note when `RWS` takes first place, it only improves over our approach marginally, however, when it underperforms, it often does so severely. This is not surprising as both `RFSF-DP` and `RFSF-TRP` approximate the signature kernel, which is a universal kernel on time series; it is theoretically capable of learning from any kind of time series data as supported by its best overall performance above.

5. Conclusion. We constructed a random kernel $\tilde{\mathsf{k}}_{\mathtt{Sig}_{\leq M}}$ for sequences that benefits from (i) lifting the original sequence to an infinite-dimensional `RKHS` $\mathcal{H}$, (ii) linear complexity in sequence length, (iii) being with high probability close to the signature kernel $\mathsf{k}_{\mathtt{Sig}}$. Thereby it combines the strength of the signature kernel $\mathsf{k}_{\mathtt{Sig}}$ which is to implicitly use the iterated integrals of a sequence that has an infinite-dimensional `RKHS` $\mathcal{H}$ as state-space with the strength

[8] https://github.com/tufts-ml/fNIRS-mental-workload-classifiers
[9] https://cloudstor.aarnet.edu.au/plus/index.php/s/pRLVtQyNhxDdCoM

of (unkernelized) signature features $\varphi_{\mathtt{Sig}}$ that only require linear time complexity. Our main theoretical result extends the theoretical guarantees for translation-invariant kernels on linear spaces to the signature kernel defined on the nonlinear domain $\mathcal{X}_{\mathtt{seq}}$; however, the proofs differ from the classic case and require to analyse the error propagation in tensor space. A second step is more straightforward, and combines this approach with random projections in finite-dimensions for tensors to reduce the complexity in memory further. The advantages and disadvantages of the resulting approach are analogous to the classic RFF technique on $\mathbb{R}^d$, namely a reduction of computational complexity by an order for the price of an approximation that only holds with high probability. As in the classic RFF case, our experiments indicate that this is in general a favourable tradeoff.

In the future, it would be interesting both theoretically and empirically to replace the vanilla Monte Carlo integration in the RFF construction by block-orthogonal random matrices as done in [88]. Further, our random features can also be used to define an unbiased appoximation to the inner product of expected signatures, which has found usecases, among many, in nonparametric hypothesis testing and market regime detection [11, 30], training of generative models [53, 31], and graph representation learning [78].

Acknowledgements. CT was supported by the Mathematical Institute Award by the University of Oxford. HO was supported by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA) and by the EPSRC grant Datasig [EP/S026347/1].

References.

[1] H. Andrès, A. Boumezoued, and B. Jourdain, *Signature-based validation of real-world economic scenarios*, 2023, https://arxiv.org/abs/2208.07251.

[2] H. Avron, M. Kapralov, C. Musco, C. Musco, A. Velingker, and A. Zandieh, *Random Fourier features for kernel ridge regression: Approximation bounds and statistical guarantees*, in International Conference on Machine Learning, 2017, pp. 253–262.

[3] F. Bach, *Sharp analysis of low-rank kernel matrix approximations*, in Conference on Learning Theory, 2013, pp. 185–209.

[4] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press, 2013.

[5] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, *Geometric deep learning: Grids, groups, graphs, geodesics, and gauges*, 2021, https://arxiv.org/abs/2104.13478.

[6] H. Buehler, B. Horvath, T. Lyons, I. Arribas, and B. Wood, *Generating financial markets with signatures*, 2021. https://ssrn.com/abstract=3657366.

[7] L. Carratino, A. Rudi, and L. Rosasco, *Learning with SGD and random features*, in Advances in Neural Information Processing Systems, 2018, pp. 10192–10203.

[8] T. Cass, T. Lyons, and X. Xu, *General signature kernels*, 2021, https://arxiv.org/abs/2107.00447.

[9] L. Chamakh, E. Gobet, and Z. Szabó, *Orlicz random Fourier features*, Journal of Machine Learning Research, 21 (2020), pp. 1–37.

[10] I. Chevyrev and A. Kormilitzin, *A primer on the signature method in machine learning*, 2016, https://arxiv.org/abs/1603.03788, https://arxiv.org/abs/1603.03788.

[11] I. Chevyrev and H. Oberhauser, *Signature moments to characterize laws of stochastic processes*, Journal of Machine Learning Research, 23 (2022), pp. 1–42.

[12] K. Choromanski, M. Rowland, T. Sarlos, V. Sindhwani, R. Turner, and A. Weller, *The geometry of random features*, in International Conference on Artificial Intelligence and Statistics, 2018, pp. 1–9.
[13] K. Choromanski, M. Rowland, and A. Weller, *The unreasonable effectiveness of structured random orthogonal embeddings*, in International Conference on Neural Information Processing Systems, 2017, pp. 218–227.
[14] K. Choromanski and V. Sindhwani, *Recycling randomness with structure for sublinear time kernel expansions*, in International Conference on Machine Learning, 2016, pp. 2502–2510.
[15] K. M. Choromanski, H. Lin, H. Chen, A. Sehanobish, Y. Ma, D. Jain, J. Varley, A. Zeng, M. S. Ryoo, V. Likhosherstov, D. Kalashnikov, V. Sindhwani, and A. Weller, *Hybrid random features*, in International Conference on Learning Representations, 2022, https://openreview.net/forum?id=EMigfE6ZeS.
[16] C. Cuchiero, L. Gonon, L. Grigoryeva, J.-P. Ortega, and J. Teichmann, *Discrete-time signatures and randomness in reservoir computing*, IEEE Transactions on Neural Networks and Learning Systems, 33 (2022), pp. 6321–6330.
[17] M. Cuturi, *Fast global alignment kernels*, in International Conference on Machine Learning, 2011, pp. 929–936.
[18] H. A. Dau, A. Bagnall, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, and E. Keogh, *The UCR time series archive*, IEEE/CAA Journal of Automatica Sinica, 6 (2019), pp. 1293–1305.
[19] J. Demšar, *Statistical comparisons of classifiers over multiple data sets*, Journal of Machine Learning Research, 7 (2006), pp. 1–30.
[20] J. Diehl, K. Ebrahimi-Fard, and N. Tapia, *Generalized iterated-sums signatures*, Journal of Algebra, 632 (2023), p. 801–824.
[21] J. Dyer, P. Cannon, and S. M. Schmon, *Approximate Bayesian computation with path signatures*, 2023, https://arxiv.org/abs/2106.12555.
[22] J. Dyer, P. W. Cannon, and S. M. Schmon, *Amortised likelihood-free inference for expensive time-series simulators with signatured ratio estimation*, in International Conference on Artificial Intelligence and Statistics, 2022, pp. 11131–11144.
[23] C. Feng, Q. Hu, and S. Liao, *Random feature mapping with signed circulant matrix projection*, in International Joint Conference on Artificial Intelligence, 2015, p. 3490–3496.
[24] A. Fermanian, P. Marion, J.-P. Vert, and G. Biau, *Framing RNN as a kernel method: A neural ODE approach*, in Advances in Neural Information Processing Systems, 2021, pp. 3121–3134.
[25] M. Fliess, *Fonctionnelles causales non linéaires et indéterminées non commutatives*, Bulletin de la société mathématique de France, 109 (1981), pp. 3–40.
[26] P. K. Friz and M. Hairer, *A course on rough paths*, Springer, 2020.
[27] K. Fukumizu, A. Gretton, B. Schölkopf, and B. K. Sriperumbudur, *Characteristic kernels on groups and semigroups*, in Advances in Neural Information Processing Systems, 2008, p. 473–480.
[28] C. Giusti and D. Lee, *Signatures, lipschitz-free spaces, and paths of persistence diagrams*, 2023, https://arxiv.org/abs/2108.02727.
[29] B. Hambly and T. Lyons, *Uniqueness for the signature of a path of bounded variation*

and the reduced path group, Annals of Mathematics, 171 (2010), pp. 109–167.
[30] B. Horvath and Z. Issa, *Non-parametric online market regime detection and regime clustering for multidimensional and path-dependent data structures*, 2023. https://ssrn.com/abstract=4493344.
[31] Z. Issa, B. Horvath, M. Lemercier, and C. Salvi, *Non-adversarial training of neural SDEs with signature kernel scores*, 2023, https://arxiv.org/abs/2305.16274.
[32] S. Janson, *Gaussian Hilbert Spaces*, Cambridge University Press, 1997.
[33] W. B. Johnson, J. Lindenstrauss, and G. Schechtman, *Extensions of Lipschitz maps into Banach spaces*, Israel Journal of Mathematics, 54 (1986), pp. 129–138.
[34] P. Kidger, J. Foster, X. Li, H. Oberhauser, and T. J. Lyons, *Neural SDEs as infinite-dimensional GANs*, in International Conference on Machine Learning, 2021, pp. 5453–5463.
[35] P. Kidger and T. Lyons, *Signatory: differentiable computations of the signature and logsignature transforms, on both CPU and GPU*, in International Conference on Learning Representations, 2021, https://openreview.net/forum?id=lqU2cs3Zca.
[36] F. J. Király and H. Oberhauser, *Kernels for sequentially ordered data*, Journal of Machine Learning Research, 20 (2019), pp. 1–45.
[37] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM review, 51 (2009), pp. 455–500.
[38] S. Kpotufe and B. Sriperumbudur, *Gaussian sketching yields a JL lemma in RKHS*, in International Conference on Artificial Intelligence and Statistics, 2020, pp. 3928–3937.
[39] S. Lang, *Algebra*, Springer, 2002.
[40] S. Lanthaler and N. H. Nelsen, *Error bounds for learning with vector-valued random features*, in Advances in Neural Information Processing Systems, 2023, https://openreview.net/forum?id=sLr1sohnmo.
[41] Q. Le, T. Sarlós, A. Smola, et al., *Fastfood-approximating kernel expansions in loglinear time*, in International Conference on Machine Learning, 2013, p. 8.
[42] D. Lee and R. Ghrist, *Path signatures on Lie groups*, 2020, https://arxiv.org/abs/2007.06633.
[43] D. Lee and H. Oberhauser, *The signature kernel*, 2023, https://arxiv.org/abs/2305.04625.
[44] M. Lemercier, C. Salvi, T. Cass, E. V. Bonilla, T. Damoulas, and T. J. Lyons, *SigGPDE: Scaling sparse Gaussian processes on sequential data*, in International Conference on Machine Learning, 2021, pp. 6233–6242.
[45] Z. Li, J.-F. Ton, D. Oglic, and D. Sejdinovic, *Towards a unified analysis of random Fourier features*, in International Conference on Machine Learning, 2019, pp. 3905–3914.
[46] Z. Liao, R. Couillet, and M. W. Mahoney, *A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent*, in Advances in Neural Information Processing Systems, 2020, pp. 13939–13950.
[47] F. Liu, X. Huang, Y. Chen, and J. A. Suykens, *Random features for kernel approximation: A survey on algorithms, theory, and beyond*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 44 (2021), pp. 7128–7148.
[48] T. Lyons, *Rough paths, signatures and the modelling of functions on streams*, 2014,

https://arxiv.org/abs/1405.4537.
[49] T. Lyons and H. Oberhauser, *Sketching the order of events*, 2017, https://arxiv.org/abs/1708.09708.
[50] T. J. Lyons, M. Caruana, and T. Lévy, *Differential equations driven by rough paths*, Springer, 2007.
[51] J. Morrill, A. Fermanian, P. Kidger, and T. Lyons, *A generalised signature method for multivariate time series feature extraction*, 2021, https://arxiv.org/abs/2006.00873.
[52] F. J. Murray and J. v. Neumann, *On rings of operators*, Annals of Mathematics, (1936), pp. 116–229.
[53] H. Ni, L. Szpruch, M. Sabate-Vidales, B. Xiao, M. Wiese, and S. Liao, *Sig-Wasserstein GANs for time series generation*, in International Conference on AI in Finance, 2022, pp. 1–8.
[54] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research, 12 (2011), pp. 2825–2830.
[55] F. Petitjean, J. Inglada, and P. Gançarski, *Satellite image time series analysis under time warping*, IEEE transactions on geoscience and remote sensing, 50 (2012), pp. 3081–3095.
[56] A. Rahimi and B. Recht, *Random features for large-scale kernel machines*, in Advances in Neural Information Processing Systems, 2007, pp. 1177–1184.
[57] A. Rahimi and B. Recht, *Uniform approximation of functions with random bases*, in Allerton Conference on Communication, Control, and Computing, 2008, pp. 555–561.
[58] A. Rahimi and B. Recht, *Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning*, Advances in Neural Information Processing Systems, (2008), pp. 1313–1320.
[59] B. Rakhshan and G. Rabusseau, *Tensorized random projections*, in International Conference on Artificial Intelligence and Statistics, 2020, pp. 3306–3316.
[60] S. Raschka, J. Patterson, and C. Nolet, *Machine learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence*, Information, 11 (2020).
[61] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.
[62] C. Reutenauer, *Free Lie algebras*, Handbook of Algebra, 3 (2003), pp. 887–903.
[63] A. Rudi and L. Rosasco, *Generalization properties of learning with random features*, in Advances in Neural Information Processing Systems, 2017, pp. 3218–3228.
[64] W. Rudin, *Fourier Analysis on Groups*, Dover Publications, 2017.
[65] C. Salvi, T. Cass, J. Foster, T. Lyons, and W. Yang, *The signature kernel is the solution of a Goursat PDE*, SIAM Journal on Mathematics of Data Science, 3 (2021), pp. 873–899.
[66] B. Schölkopf and A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, 2002.
[67] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge

University Press, 2004.
[68] B. Sriperumbudur, K. Fukumizu, and G. Lanckriet, *On the relation between universality, characteristic kernels and RKHS embedding of measures*, in International Conference on Artificial Intelligence and Statistics, 2010, pp. 773–780.
[69] B. Sriperumbudur and Z. Szabó, *Optimal rates for random Fourier features*, in Advances in Neural Information Processing Systems, 2015, pp. 1144–1152.
[70] B. K. Sriperumbudur and N. Sterge, *Approximate kernel PCA using random features: Computational vs. statistical trade-off*, Annals of Statistics, (2022), pp. 2713–2736.
[71] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer Science & Business Media, 2008.
[72] Y. Sun, A. Gilbert, and A. Tewari, *But how does it work in theory? Linear SVM with random features*, in Advances in Neural Information Processing Systems, 2018, pp. 3379–3388.
[73] Y. Sun, Y. Guo, J. A. Tropp, and M. Udell, *Tensor random projection for low memory dimension reduction*, 2021, https://arxiv.org/abs/2105.00105.
[74] D. J. Sutherland and J. Schneider, *On the error of random Fourier features*, in Conference on Uncertainty in Artificial Intelligence, 2015, pp. 862–871.
[75] Z. Szabó and B. Sriperumbudur, *On kernel derivative approximation with random Fourier features*, in International Conference on Artificial Intelligence and Statistics, 2019, pp. 827–836.
[76] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, *Fourier features let networks learn high frequency functions in low dimensional domains*, in Advances in Neural Information Processing Systems, 2020, pp. 7537–7547.
[77] C. Toth, P. Bonnier, and H. Oberhauser, *Seq2Tens: An efficient representation of sequences by low-rank tensor projections*, in International Conference on Learning Representations, 2021, https://openreview.net/forum?id=dx4b7lm8jMM.
[78] C. Toth, D. Lee, C. Hacker, and H. Oberhauser, *Capturing graphs with hypo-elliptic diffusions*, in Advances in Neural Information Processing Systems, 2022, pp. 38803–38817.
[79] C. Tóth and H. Oberhauser, *Bayesian learning from sequential data using Gaussian processes with signature covariances*, in International Conference on Machine Learning, 2020, pp. 9548–9560.
[80] E. Ullah, P. Mianjy, T. V. Marinov, and R. Arora, *Streaming kernel PCA with $\tilde{O}(\sqrt{n})$ random features*, in Advances in Neural Information Processing Systems, 2018.
[81] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge University Press, 2018.
[82] J. Wacker and M. Filippone, *Local random feature approximations of the Gaussian kernel*, Procedia Computer Science, 207 (2022), pp. 987–996.
[83] J. Wacker, M. Kanagawa, and M. Filippone, *Improved random features for dot product kernels*, 2022, https://arxiv.org/abs/2201.08712.
[84] J. Wacker, R. Ohana, and M. Filippone, *Complex-to-real sketches for tensor products with applications to the polynomial kernel*, in International Conference on Artificial Intelligence and Statistics, 2023, pp. 5181–5212.
[85] C. Williams and M. Seeger, *Using the Nyström method to speed up kernel machines*,

in Advances in Neural Information Processing Systems, 2000, pp. 682–688.

[86] L. Wu, I. E.-H. Yen, J. Yi, F. Xu, Q. Lei, and M. Witbrock, *Random warping series: A random features method for time-series embedding*, in International Conference on Artificial Intelligence and Statistics, 2018, pp. 793–802.

[87] T. Yokonuma, *Tensor Spaces and Exterior Algebra*, American Mathematical Society, 1992.

[88] F. X. X. Yu, A. T. Suresh, K. M. Choromanski, D. N. Holtmann-Rice, and S. Kumar, *Orthogonal random features*, in Advances in Neural Information Processing Systems, 2016, pp. 1975–1983.