The Annals of Applied Probability
2024, Vol. 34, No. 1A, 585–626
https://doi.org/10.1214/23-AAP1973

WEIGHTED SIGNATURE KERNELS

BY THOMAS CASS[1,a], TERRY LYONS[2,b] AND XINGCHENG XU[3,c]

[1]*Department of Mathematics, Imperial College London,* [a]*thomas.cass@imperial.ac.uk*
[2]*Matematical Institute, University of Oxford,* [b]*t.lyons@maths.ox.ac.uk*
[3]*Department of Mathematics, Imperial College London,* [c]*xingcheng.xu18@gmail.com*

Suppose that γ and σ are two continuous bounded variation paths, which take values in a finite-dimensional inner product space V. The recent papers (*J. Mach. Learn. Res.* **20** (2019) 1–45) and (*SIAM J. Math. Data Sci.* **3** (2021) 873–899), respectively, introduced the truncated and the untruncated signature kernel of γ and σ, and showed how these concepts can be used in classification and prediction tasks involving multivariate time series. In this paper, we introduce signature kernels $K_{\phi}^{\gamma,\sigma}$ indexed by a weight function ϕ, which generalise the ordinary signature kernel. We show how $K_{\phi}^{\gamma,\sigma}$ can be interpreted in many examples as an average of PDE solutions, and thus we show how it can be estimated computationally using suitable quadrature formulae. We extend this analysis to derive closed-form formulae for expressions involving the expected (Stratonovich) signature of Brownian motion. In doing so, we articulate a novel connection between signature kernels and the notion of the hyperbolic development of a path, which has been a broadly useful tool in the recent analysis of the signature; see, for example, (*Ann. of Math. (2)* **171** (2010) 109–167; *J. Funct. Anal.* **272** (2017) 2933–2955) and (*Trans. Amer. Math. Soc.* **372** (2019) 585–614). As applications, we evaluate the use of different general signature kernels as a basis for nonparametric goodness-of-fit tests to Wiener measure on path space.

1. Introduction. Kernel methods are well-established tools in machine learning, which are fundamental to support vector machine models for classification, nonlinear regression and outlier detection involving small or moderate-sized data sets [4, 27, 29]. Applications are manifold and include text classification [20], protein classification [19] as well as applications to biological sequences [32] and labelled graphs [16]. The essence of these methods is to achieve better separation between labelled data by embedding a low-dimensional feature space X into a higher-dimensional one H, which is commonly assumed to be a Hilbert space, by means of a feature map $\psi : X \to H$. The associated kernel is a function $K : X \times X \to \mathbb{R}$ with the property that $\langle \psi(x), \psi(y)\rangle_H = K(x, y)$ for all x and y in X. If K is known in closed form, then the inner products of all extended features are obtainable from the evaluation of K at pairs of training instances in the original feature set. A typical classification problem can be formulated as convex constrained optimisation problem for which the Lagrangian dual involves only the inner products of pairs of enhanced features in the set of training instances. Crucially, one does not need the vectors of the enhanced features themselves. This observation—the basis of the so-called kernel trick—then allows one to enjoy the advantages of working in a higher-dimensional feature space without some of the concomitant drawbacks. The references [15, 25] give an extensive treatment of the core of this theory and a survey of its applications

The selection of an effective kernel is challenging and somewhat task-dependent. When the training data consist of sequential data such as time series, these challenges are magnified. To

Received July 2021; revised April 2023.
MSC2020 subject classifications. 60L10.
Key words and phrases. The signature, expected signatures, kernel methods, general signature kernels, Gaussian quadrature, hyperbolic development, contour integration.

address these and other difficulties, much recent progress has been made by re-purposing the (path) signature transform, which has decisive advantages in capturing complex interactions between multivariate data streams. The signature originates in the fundamental work of Chen [7] and [6]. We recall that for a continuous bounded variation path $\gamma : [a, b] \to V$ it is given by the formal tensor series of iterated integrals

$$(1.1)\qquad \begin{aligned} S(\gamma)_{a,b} &= 1 + \sum_{k=1}^{\infty} S(\gamma)_{a,b}^{k} \in T((V)) = \prod_{k=0}^{\infty} V^{\otimes k} \quad \text{with} \\ S(\gamma)_{a,b}^{k} &:= \int_{a<t_1<t_2<\cdots<t_k<b} d\gamma_{t_1} \cdots d\gamma_{t_k}. \end{aligned}$$

The soundness of capturing γ through $S(\gamma)$ is underpinned by the fact that the map $\gamma \mapsto S(\gamma)_{a,b}$ is one-to-one, up to an equivalence relation on the space of paths [14]. The signature is invariant under reparameterisation and, therefore, by representing the path γ by the tensor series $S(\gamma)_{a,b}$ one removes an otherwise complicating infinite-dimensional symmetry. On the other hand, the signature captures the order of events along γ. The algebraic properties of the signature have been developed since the foundational work of Chen; it is now well understood that the signature transform describes the set of polynomials on unparameterised paths, in a sense that can be made meaningful. Analytically, the signature of γ characterises the class of responses (i.e., solutions) of all smooth differential systems, which have γ as the input.

An important fact is the factorial decay rate of the terms in the series in (1.1). That is, given appropriately defined norms on the tensor product spaces $V^{\otimes k}$:

$$\left\| \int_{a<t_1<t_2<\cdots<t_k<b} d\gamma_{t_1} \cdots d\gamma_{t_k} \right\|_{V^{\otimes k}} \le \frac{L(\gamma)^k}{k!},$$

where $L(\gamma)$ denotes the length of the path over $[a, b]$. This allows one to define the (untruncated) signature kernel of two paths γ and σ by

$$(1.2)\qquad K_{s,t}(\gamma, \sigma) = \langle S(\gamma)_{a,s}, S(\sigma)_{a,t} \rangle := 1 + \sum_{k=1}^{\infty} \langle S(\gamma)_{a,s}^{k}, S(\sigma)_{a,t}^{k} \rangle_k,$$

where $\langle \cdot, \cdot \rangle_k$ is the canonical (Hilbert–Schmidt) inner product on $V^{\otimes k}$ derived from a fixed inner product on V. In the recent paper [23], it was shown that this untruncated signature kernel has some advantages over it truncated counterpart [17], which in some cases, lead to greater accuracy in classification and regression tasks on benchmark data sets for multivariate time series. The explanation for this turns on the key observation that for continuous paths of bounded variation with almost everywhere defined derivatives, K is the unique solution of the hyperbolic partial differential equation

$$(1.3)\qquad \frac{\partial^2 K}{\partial s \partial t}(s, t) = K(s, t) \langle \gamma_s', \sigma_t' \rangle \quad \text{with} \quad K(a, \cdot) = K(\cdot, a) \equiv 1.$$

The solution to which can be approximated using PDE solvers, thus allowing for the efficient computation of the inner product in (1.2) and obviating the need to compute iterated integrals.

While the kernel (1.2) is useful, it is also in some respects confining. One restriction it imposes is on the relative contributions made to the sum (1.2) by the different inner products $\langle \cdot, \cdot \rangle_k$. It is easy to see, for example, by scaling γ by $\lambda = e^{\alpha} \in \mathbb{R}$ to give $(\lambda\gamma)_\cdot = \lambda\gamma_\cdot$, we have

$$K^{\lambda\gamma,\sigma}(s, t) = 1 + \sum_{k=1}^{\infty} e^{\alpha k} \langle S(\gamma)_{a,s}^{k}, S(\sigma)_{a,t}^{k} \rangle_k,$$

so that the signature kernel for the family of inner products $\langle \cdot, \cdot \rangle_\alpha = \sum_{k \geq 0} e^{\alpha k} \langle \cdot, \cdot \rangle_k$ can be obtained as above by solving the appropriately rescaled version of the PDE (1.3). The starting point for this paper is to introduce methods that allow for the efficient computation of general signature kernels with a broad class of different weightings. These will be derived from bilinear forms on $T(V)$ of the type

$$\langle \cdot, \cdot \rangle_\phi = \sum_{k=0}^{\infty} \phi(k) \langle \cdot, \cdot \rangle_k,$$

where $\phi : \mathbb{N} \cup \{0\} \to \mathbb{R}$ (or, sometimes, $\mathbb{C}$), so that $\langle \cdot, \cdot \rangle_\phi$ need not even define an inner product. For a given such ϕ, we term the resulting kernel the ϕ-signature kernel. One fundamental observation we take advantage of is illustrated by the following argument: assume $\phi(0) = 1$, and suppose that we can solve the Hamburger moment problem for the sequence $\{\phi(k) : k \in \mathbb{N} \cup \{0\}\}$, that is, we can find a probability measure μ on $\mathbb{R}$ such that

$$\phi(k) = \int \lambda^k \, d\mu(\lambda) \quad \text{for all } k \in \mathbb{N} \cup \{0\}. \tag{1.4}$$

Then, under some conditions on μ, we will be able to justify the following identity:

$$\langle S(\gamma)_{a,s}, S(\sigma)_{a,t} \rangle_\phi = \sum_{k=0}^{\infty} \int \lambda^k \langle S(\gamma)_{a,s}^k, S(\sigma)_{a,t}^k \rangle_k \, d\mu(\lambda) = \int K^{\lambda\gamma,\sigma}(s,t) \, d\mu(\lambda). \tag{1.5}$$

In this case, the computation of the ϕ-signature kernel, that is, the one arising from $\langle \cdot, \cdot \rangle_\phi$, will amount to integrating scaled solutions to the PDE (1.3) in λ with respect to the measure μ. The practicability of this approach depends on two aspects. First, one needs to be able to solve the moment problem (1.4); there are well-known necessary and sufficient conditions but, ideally, μ should be determined explicitly. Second, one needs to be able to approximate accurately the integral on the right-hand side of (1.5). In this respect, one is helped by the form of the function $\lambda \mapsto K^{\lambda\gamma,\sigma}(s,t)$, which is real analytic with a power series whose coefficients decay at rate $(n!)^{-2}$. Hence, in cases where μ has a density w given in closed form, Gaussian quadrature provides an approximation of the form

$$\int K^{\lambda\gamma,\sigma}(s,t) \, d\mu(\lambda) \approx \sum_{i=1}^{m} w_i K^{\lambda_i\gamma,\sigma}(s,t)$$

and equip us with well-described error bounds; see, for example, [30]. For these examples, the ϕ-signature kernel can be approximated at the expense of m implementations of a PDE solver.

The same principle outlined in the previous paragraph can appear in different guises. For example, by solving the trigonometric moment problem

$$\phi(k) = \int_0^{2\pi} e^{ik\theta} \, d\mu(\theta) \quad \text{for } k \in \mathbb{Z}$$

to find a measure μ on $[0, 2\pi]$, then an analogue of (1.5) can be obtained by integrating the complex-valued function $\theta \mapsto K^{\exp(i\theta)\gamma,\sigma}(s,t)$ with respect to μ. A similar observation applies to a class of integral transforms having the form

$$\phi(u) = \int_C r(u,z) \, d\mu(z) \quad \text{where } r(u,z) = g(z)^{\alpha u} \in \mathbb{C} \text{ for some } \alpha \in \mathbb{R}. \tag{1.6}$$

This class includes the Fourier–, Laplace– and Mellin–Stieltjes transforms, for which specific pairs (ϕ, μ) are of course extensively documented. We illustrate a range of examples that can be generated using this idea in the main text.

Extensions of the same idea apply to expected signatures. It is by now well known that, under some conditions, the expected signature of a stochastic process characterises the law of that process [8]. This motivates the use of expected signatures as a measure of similarity of two laws on path space, for example, through the quantity

$$d_\phi(\mu,\nu) := \|\mathbb{E}^\mu[S(X)] - \mathbb{E}^\nu[S(X)]\|_\phi,$$

which is seen to be a maximum mean discrepancy (MMD) distance between μ and ν; see [13] and [9]. We also have a measure of alignment of the two expected signatures of μ and ν given by

$$\cos\angle_\phi(\mu,\nu) := \frac{\langle \mathbb{E}^\mu[S(X)], \mathbb{E}^\nu[S(X)]\rangle_\phi}{\|\mathbb{E}^\mu[S(X)]\|_\phi \|\mathbb{E}^\nu[S(X)]\|_\phi},$$

which can be interpreted as an analogue of the Pearson correlation coefficient for measures on path space. As an application, we consider designing goodness-of-fit tests in which one wants to understand when an observed empirical sample is drawn from a well-described baseline distribution. A motivating example for this paper was that of the detection of radio frequency interference (RFI) contamination in radio astronomy. In this situation, electrical signals are collected from an array of antennas [34]. Under the null hypothesis of no RFI contamination, the signals will reflect only the so-called thermal noise of the receiving equipment. From this perspective, the most important reference distribution is that of white noise or, in its integrated form, Brownian motion. Kernels have been used for similar problems previously, albeit for the case of vector-valued data; see, for example, [10]. Proposals have been made to put similar ideas into practice in the context of two-sample statistical tests to determine whether two observed empirical measures on paths are drawn from the same underlying distribution; see, for example, [9] and [18].

A formula for the expected Stratonovich signature of multivariate Brownian motion has been known since the work of Fawcett [12] and Lyons and Victoir [21]. In the context of the problems described above, we can take advantage of Fawcett's formula to prove what we believe to be a novel identity, namely that for any continuous path γ of bounded variation, we have

$$\big\langle \mathbb{E}[S(\circ B)_{0,s}], S(\gamma)_{0,t}\big\rangle_\phi = \cosh\big(\rho_{\sqrt{s/2}\gamma}(t)\big). \tag{1.7}$$

In this formula, $\rho_\gamma(t)$ is the hyperbolic distance between the starting point and the end point of the hyperbolic development of the path segment $.\gamma|_{[0,t]}$, and

$$\phi(k) := \Gamma(k/2+1) := \int_0^\infty x^{k/2}e^{-x}\,dx.$$

When we realise hyperbolic space as a hyperboloid, the right-hand side of formula (1.7) can be obtained by solving a linear ordinary differential equation. In the special case where γ is piecewise linear, this solution of the equation is a known product of matrices. These remarks allow one to compute quantities like $d_\phi(\mathcal{W},\nu)$, where $\mathcal{W}$ denotes Wiener measure and ν is an empirical measure on bounded variation paths. We note that the primary use of the hyperbolic development in the study of signatures to date has been in obtaining lower bounds for the study of signature asymptotics; see [14] and [2]. In this context, the identity (1.7) appears new, and it establishes a connection between the signature kernel and these broader topics. It seems plausible that an additional benefit of (1.7) will be that it allows a more analytic treatment of these other problems in a way that relies less on the geometrical intricacies of hyperbolic space.

If $\phi\equiv 1$, we can use Hankel's well-known representation for the reciprocal Gamma function as the contour integral

$$\frac{1}{\Gamma(z)} = \frac{1}{2\pi i}\oint_H w^{-z}e^w\,dw,$$

where H is Hankel's contour. Noting the similarity with (1.6) we can obtain the identity

$$\langle \mathbb{E}[S(\circ B)_{0,s}], S(\gamma)_{0,t}\rangle_\phi = \frac{1}{2\pi i}\oint_C w^{-1}e^w \cosh\bigl(\rho_{\sqrt{s/2w}\gamma}(t)\bigr)\,dw, \tag{1.8}$$

for an appropriate contour C. To make sense of this formula, we first need to make sense of the complex rescaling in the defining ODE for hyperbolic development. The numerical evaluation of contour integrals of the form $\oint_C f(w)e^w\,dw$ is an active topic in numerical integration (see [24]), and we use these ideas to evaluate (1.8). The same idea can be extended to cover general ϕ.

In the final two sections, we consider examples, which lend themselves to being treated by the methods outlined above. A natural question is how to select an appropriate ϕ for a given task, and the related question of how to evaluate the performance of a given kernel against an alternative. To develop this, we reverse the perspective taken above and use d_ϕ to define a loss function

$$L_\phi(\mathcal{W}, \mu) := d_\phi(\mathcal{W}, \mu)^2,$$

and given a finite collection of paths $\{\gamma_1, \dots, \gamma_n\}$, we consider the problem of minimising L over the set

$$C_n = \left\{\mu = \sum_{i=1}^n \lambda_i \delta_{\gamma_i} : \sum_{i=1}^n \lambda_i = 1, \lambda_i \geq 0\right\}.$$

Under some conditions on the support, this optimisation problem will have a unique solution μ^* which we can find. This gives us a way of evaluating the similarity of a given finitely supported (possibly empirical) measure μ to Wiener measure under the loss function induced by $\langle\cdot,\cdot\rangle_\phi$ by comparing $L_\phi(\mathcal{W}, \mu)$ and $L_\phi(\mathcal{W}, \mu^*)$. For example, if the ratio $\frac{L_\phi(\mathcal{W},\mu^*)}{L_\phi(\mathcal{W},\mu)} < \alpha \ll 1$, then by an appropriate selection of the threshold α, one might decide that μ does not resemble Wiener measure. We do not give an extensive treatment of examples, but to illustrate how these methods introduced above might be used we consider two cases in detail:

(1) Cubature measures of degree N on Wiener space are finitely supported measures, which matched the expected iterated integrals of Brownian motion up to and including degree N. Explicit constructions are known in some cases; see [21]. By definition, these measures will be optimal in the above sense for any kernel induced by any ϕ with $\phi(k) = 0$ for $k \geq N$. One might expect that they are close to optimal for smoother ϕ, which still decay sufficiently fast.

(2) We model radio frequency interference in sky-subtracted visibilities radio astronomy as advocated by [34] and consider two idealised types of signal contamination:

- Narrow-band RFI measure across n antennas. In this case, the received signals are n linear superpositions of independent Brownian motions with a single-frequency sinusoidal wave of a fixed amplitude.
- Short duration high energy bursts. As a model for this, we consider the gerneralisation to the multivariate case of the example, originally considered in the univariate setting in which the signal is given by $X_t = W_t + \epsilon\sqrt{(t-U)_+}$ for $t \in [0, 1]$, where $(W_t)_{t\in[0,1]}$ is a Brownian motion, U is independent an uniformly distributed on $[0, 1]$ and $\epsilon > 0$. The theoretical interest in this comes from the existence of a critical parameter $\epsilon_0 > 0$ for which the law of X is equivalent to $\mathcal{W}$ if and only if $\epsilon < \epsilon_0$ (see [11]), and which therefore gives an example that falls outside the scope of traditional maximum-likelihood-based approaches to the problem.

2. Background on general signature kernels. Let $T(V)$ denote the algebra of tensor polynomials over a finite-dimensional vector space V, which consists of elements of the form

$$a = \sum_{k=0}^{\infty} a_k, \quad a_k \in V^{\otimes k} \text{ such that } a_k = 0 \text{ for all but finitely many } k,$$

with the tensor product defined by

$$ab = \sum_{k=0}^{\infty} \sum_{l=0}^{k} a_l b_{k-l},$$

where the product $V^{\otimes l} \times V^{\otimes (k-l)} \ni (c,d) \mapsto cd \in V^{\otimes k}$ is determined by $((v_1 \cdots v_l), (v_{l+1} \cdots v_k)) \mapsto v_1 \cdots v_k$ for $v_1, \ldots, v_k \in V$. We let $T((V))$ denote the space of formal tensor series, that is, those of the form

$$a = \sum_{k=0}^{\infty} a_k \quad \text{for } a_k \in V^{\otimes k}.$$

We let V^* denote the dual space of V. Then $T(V^*)$ is the dual space of $T((V))$, and the signature of a continuous bounded variation path $\gamma : [a,b] \to V$ is the family of elements $\{S(\gamma)_{s,t} : s \leq t \in [a,b]\}$ in $T((V))$ determined by the solution to the differential equation

$$dS(\gamma)_{s,u} = S(\gamma)_{s,u} d\gamma_u \quad \text{on } [s,t] \text{ with } S_{s,s} = 1. \tag{2.1}$$

Alternatively we can view the signature as the series of iterated integrals

$$S(\gamma)_{s,t} = 1 + \sum_{k=1}^{\infty} \int_{s<t_1<\cdots<t_k<t} d\gamma_{t_1} \cdots d\gamma_{t_k} \in T((V)). \tag{2.2}$$

We denote the range of the signature map by

$$\mathcal{S} = \{S(\gamma)_{s,t} : \gamma, s < t\} \subset T((V)). \tag{2.3}$$

We consider dual pairs (E, F), where E and F are two linear subspaces of $T((V))$. Recall that this means that $(\cdot,\cdot) : E \times F \to \mathbb{R}$ is a bilinear map such that the linear functionals $\{(e,\cdot) : e \in E\} \subset F^*$ and $\{(\cdot, f) : f \in F\} \subset E^*$ separate points in F and E, respectively. We thus can identify E and F as linear subspaces of the algebraic dual spaces F^* and E^*, respectively.

DEFINITION 2.1. Let (E, F) be a dual pair. Suppose that $\mathcal{S} \subset E \cap F$ where $\mathcal{S}$ denotes the range of the signature map (2.3). Then given two continuous paths $\gamma, \sigma : [a,b] \to V$ of bounded variation, we define the $(\cdot,\cdot)$-signature kernel of γ and σ to be the function

$$[a,b] \times [a,b] \ni (s,t) \mapsto (S(\gamma)_{a,s}, S(\sigma)_{a,t}) = K^{\gamma,\sigma}_{(\cdot,\cdot)}(s,t).$$

REMARK 2.2. This definition is not symmetric in general, that is, it may hold that $K^{\gamma,\sigma}_{(\cdot,\cdot)} \neq K^{\sigma,\gamma}_{(\cdot,\cdot)}$.

For this definition to be useful, we need to demand more of the pairing (E, F). More exactly, we need at least that their continuous duals satisfy $F \subseteq E'$ and $E \subseteq F'$. To go further still, we will need that they respect some of the algebraic structure on $T((V))$. The examples we will work are derived from a fixed but arbitrary inner product $\langle \cdot,\cdot \rangle$ on V. This gives rise to the Hilbert–Schmidt inner product $\langle \cdot,\cdot \rangle_k$ on the k-fold tensor product spaces $V^{\otimes k}$ in a canonical way. Then, by taking

$$\langle a,b \rangle_\phi := \sum_{k=0}^{\infty} \phi(k) \langle a_k, b_k \rangle_k$$

for some weight function $\phi : \mathbb{N} \cup \{0\} \to \mathbb{R}_+$ we may define $T_\phi((V))$ to be the Hilbert space obtained by completing $T(V)$ with respect to $\langle\cdot,\cdot\rangle_\phi$. We equip $T_\phi((V))$ with the norm topology unless stated otherwise. It is necessary to have a condition on ϕ, which ensures that $\mathcal{S} \subset T_\phi((V))$.

LEMMA 2.3. *Let $\phi : \mathbb{N} \cup \{0\} \to \mathbb{R}_+$ be such that for every $C > 0$ the series $\sum_{k\in\mathbb{N}} C^k \phi(k)(k!)^{-2}$ is summable. Then $\mathcal{S} \subset T_\phi((V))$.*

PROOF. Let $\{e_i : i = 1, \dots, d\}$ be any orthonormal basis of V w.r.t. $\langle\cdot,\cdot\rangle$, and $\{e_I^* : I = (i_1, \dots, i_k)\}$ the associated dual basis on $(V^*)^{\otimes k}$. Then

$$\|S(\gamma)_{s,t}\|_\phi^2 = \sum_{k=0}^{\infty} \phi(k) \sum_{|I|=k} \left[S(\gamma)_{s,t}(e_I^*)\right]^2, \tag{2.4}$$

where

$$S(\gamma)_{s,t}(e_I^*) := S^I(\gamma)_{s,t} = \int_{s<u_1<u_2<\cdots<u_k<t} d\langle e_{i_1}, \gamma_{u_1}\rangle d\langle e_{i_2}, \gamma_{u_2}\rangle \cdots d\langle e_{i_k}, \gamma_u\rangle.$$

We estimate the summands in (2.4) by

$$\sum_{|I|=k} \left[S^I(\gamma)_{s,t}\right]^2 \le \frac{L_{s,t}(\gamma)^{2k}}{(k!)^2} \quad \text{where } L_{s,t}(\gamma) := \int_s^t |d\gamma_u| \text{ is the length of } \gamma.$$

The summability condition then ensures that (2.4) is finite. □

REMARK 2.4. The summability condition is also necessary in order that $\mathcal{S} \subset T_\phi((V))$ because $\mathcal{S}$ contains paths of the form $\gamma_t = tv$ for arbitrary $v \in V$ for which

$$\|S(\gamma)_{s,t}\|_\phi^2 = \sum_{k=0}^{\infty} \phi(k)(k!)^{-2}\|v\|^2(t-s)^{2k}.$$

This prompts the following condition.

CONDITION 1. The function $\phi : \mathbb{N} \cup \{0\} \to \mathbb{R}_+$ is such that the series $\sum_{k\in\mathbb{N}} C^k \phi(k) \times (k!)^{-2}$ is summable for every $C > 0$.

The next lemma describes examples of dual pairs (E, F) of Hilbert spaces, which fulfill the conditions in Definition 2.1.

LEMMA 2.5. *Let $\phi : \mathbb{N} \cup \{0\} \to \mathbb{R}_+$ and $\psi : \mathbb{N} \cup \{0\} \to \mathbb{R}_+$ be functions such that ϕ and ψ^{-1} (i.e., $n \mapsto \psi(n)^{-1}$) satisfy the summability criterion of Condition 1. In each of the following cases, (E, F) is a dual pair, which satisfies $F \subseteq E'$ and $E \subseteq F'$:*

(1) $E = T_\phi((V))$, $F = T_\phi((V))$, $(\cdot,\cdot) = \langle\cdot,\cdot\rangle_\phi$,
(2) $E = T_\phi((V))$, $F = T_{\psi^{-1}}((V))$, $(\cdot,\cdot) = \langle\cdot,\cdot\rangle_{\sqrt{\phi/\psi}}$.

PROOF. For notational ease, we write H_ϕ for $T_\phi((V))$. In both cases, Condition 1 ensures that $\mathcal{S} \subset E \cap F$. In case 1, it is classical that $H_\phi' = \{\langle h,\cdot\rangle_\phi : h \in H_\phi\}$, while for case 2 we have for $h \in H_\phi$ and $g \in H_{\psi^{-1}}$ that

$$\left|\langle h, g\rangle_{\sqrt{\phi/\psi}}\right| = \left|\sum_{k=0}^{\infty} \sqrt{\frac{\phi(k)}{\psi(k)}} \langle h_k, g_k\rangle_k\right| \le \|h\|_\phi \|g\|_{\psi^{-1}},$$

hence $\{\langle h,\cdot\rangle_{\sqrt{\phi/\psi}} : h \in H_\phi\} \subseteq H'_{\psi^{-1}}$. By using the fact that h is in H_ϕ if and only if $\tilde{h} := \sqrt{\phi\psi}h := \sum_k \sqrt{\phi(k)\psi(k)}h_k$ is in $H_{\psi^{-1}}$, we see that

$$\langle h,\cdot\rangle_{\sqrt{\phi/\psi}} = \langle \tilde{h},\cdot\rangle_{\psi^{-1}}$$

so that $\{\langle h,\cdot\rangle_{\sqrt{\phi/\psi}} : h \in H_\phi\} = \{\langle h,\cdot\rangle_{\psi^{-1}} : h \in H_{\psi^{-1}}\} = H'_{\psi^{-1}}$. $\square$

Hereafter, we will work entirely in the case $\langle T_\phi((V)), T_\phi((V))\rangle_\phi$ in which the dual pair is the Hilbert space $T_\phi((V))$ with itself with pairing given by the inner product $\langle\cdot,\cdot\rangle_\phi$. This leads to the following definition.

DEFINITION 2.6. Let $\phi : \mathbb{N}\cup\{0\} \to \mathbb{R}_+$ satisfy Condition 1. Given two continuous paths $\gamma,\sigma : [a,b] \to V$ of bounded variation, we define the *ϕ-signature kernel* of γ and σ to be the two-parameter function $K_\phi^{\gamma,\sigma}$ defined by

$$[a,b]\times[a,b] \ni (s,t) \mapsto \langle S(\gamma)_{a,s}, S(\sigma)_{a,t}\rangle_\phi =: K_\phi^{\gamma,\sigma}(s,t).$$

REMARK 2.7. It is straightforward to extend the discussion above to consider general bilinear forms of signatures. If $\phi : \mathbb{N}\cup\{0\} \to \mathbb{R}$, then we can define a semi-definite inner product on $T(V)$ by

$$\langle a,b\rangle_{|\phi|} := \sum_{k=0}^{\infty} |\phi(k)| \langle a_k, b_k\rangle_k.$$

Let N denote the linear subspace of $T(V)$ given by the kernel of semi-norm $\|\cdot\|_{|\phi|}$. Then we we can complete the quotient space $T(V)/N$ with respect to inner product $\langle\cdot,\cdot\rangle_{|\phi|}$ and denote the resulting Hilbert space by $T_{|\phi|}((V))$. By construction, the corresponding kernel will not depend on those terms in the signature whose levels coincide with the zeros of ϕ. The bilinear form on $T(V)$

$$\langle a,b\rangle_\phi := B_\phi(a,b) := \sum_{k=0}^{\infty} \phi(k)\langle a_k,b_k\rangle_k \tag{2.5}$$

extends to a continuous bilinear form on $T_{|\phi|}((V))$. If ϕ is such that $|\phi|$ satisfies Condition 1 then, as above, we define the *ϕ-signature kernel* of γ and σ to be the function $K_\phi^{\gamma,\sigma} : [a,b]\times[a,b] \to \mathbb{R}$ by

$$K_\phi^{\gamma,\sigma}(s,t) := \langle S(\gamma)_{a,s}, S(\sigma)_{a,t}\rangle_\phi.$$

This agrees with the previous definition whenever ϕ takes positive values.

The following shifted weight functions arise naturally when doing calculus on signature kernels.

DEFINITION 2.8. Given a function $\phi : \mathbb{N}\cup\{0\} \to \mathbb{R}$ and $k \in \mathbb{N}$, we define the k-shift of ϕ to be the function $\phi_{+k} : \mathbb{N}\cup\{0\} \to \mathbb{R}$ determined by $\phi_{+k}(\cdot) = \phi(\cdot + k)$.

The next result is fundamental.

PROPOSITION 2.9. *Let $\gamma,\sigma : [a,b] \to V$ be two continuous paths of bounded variation. Assume that the function $\phi : \mathbb{N}\cup\{0\} \to \mathbb{R}$ is such that $|\phi|$ and its 1-shift $|\phi_{+1}|$ both satisfy Condition 1. Then the ϕ- and ϕ_{+1}-signature kernels of γ and σ are well-defined and are related by the two-parameter integral equation*

$$K_\phi^{\gamma,\sigma}(s,t) = \phi(0) + \int_a^s \int_a^t K_{\phi_{+1}}^{\gamma,\sigma}(u,v)\langle d\gamma_u, d\sigma_v\rangle.$$

PROOF. Well-definedness of the two signature kernels follows from the summability conditions. Unravelling the definitions and using (2.1) gives

$$\begin{aligned}
K_\phi^{\gamma,\sigma}(s,t) &= \sum_{k=0}^{\infty}\phi(k)\sum_{|I|=k} S^I(\gamma)_{a,s}S^I(\sigma)_{a,t} \\
&= \phi(0) + \sum_{k=1}^{\infty}\phi(k)\sum_{|I|=k-1}\int_a^s\int_a^t S^I(\gamma)_{a,u}S^I(\sigma)_{a,v}\langle d\gamma_u, d\sigma_v\rangle \\
&= \phi(0) + \int_a^s\int_a^t \sum_{k=0}^{\infty}\phi(k+1)\sum_{|I|=k} S^I(\gamma)_{a,u}S^I(\sigma)_{a,v}\langle d\gamma_u, d\sigma_v\rangle \\
&= \phi(0) + \int_a^s\int_a^t K_{\phi_{+1}}^{\gamma,\sigma}(s,t)\langle d\gamma_u, d\sigma_v\rangle. \qquad \square
\end{aligned}$$

In the special case where ϕ is constant, we see that the shift $\phi_{+k} = \phi$ for every k and, therefore, $K_\phi^{\gamma,\sigma}$ satisfies

$$K_\phi^{\gamma,\sigma}(s,t) = \phi(0) + \int_a^s\int_a^t K_\phi^{\gamma,\sigma}(u,v)\langle d\gamma_u, d\sigma_v\rangle,$$

and in particular, if γ and σ are differentiable and $\phi \equiv 1$, then we write $K_\phi^{\gamma,\sigma} = K^{\gamma,\sigma}$ and refer to it as the *original signature kernel*. As was first shown in [23], it solves the partial differential equation

$$\frac{\partial^2 K^{\gamma,\sigma}(s,t)}{\partial s\partial t} = K^{\gamma,\sigma}(s,t)\langle\gamma_s', \sigma_t'\rangle \quad \text{on } [a,b]\times[a,b] \tag{2.6}$$

with boundary conditions $K(a,\cdot) \equiv K(\cdot,a) \equiv 1$. The same paper shows how the solution to (2.6) can be approximated numerically, and how the methodology extends to the case of rough paths. The approximate solution can then be used to implement kernel learning methods for classification or regression tasks based on time series as mentioned in the Introduction; see [9, 17].

It is self-evident from Proposition 2.9 that for general ϕ the function will not solve a PDE of the type (2.6). Nevertheless, we can produce examples of different ϕ, which do by varying the inner product $\langle\cdot,\cdot\rangle$ on the underlying vector space V, or by scaling the inner product on $T(V)$ homogeneously with respect the grading on $T(V)$. By the latter idea, we mean that, for $\theta \in \mathbb{R}$ we can define $\delta_\theta : T(V) \to T(V)$ to be the unique algebra homomorphism, which is determined by scalar multiplication by θ on V (i.e., $V \ni a \mapsto \theta a$), then we have

$$\delta_\theta a = \sum_{k=0}^{\infty}\theta^k a_k \quad \text{if } a = \sum_{k=0}^{\infty} a_k \in T(V). \tag{2.7}$$

This map can be defined in the same way on $T((V))$. The following lemma explores the properties of δ_θ as a map between Hilbert spaces of the form $T_\phi((V)) \subset T((V))$.

LEMMA 2.10. *Suppose* $0 \neq \theta \in \mathbb{R}$. *Given* $\phi : \mathbb{N}\cup\{0\} \to \mathbb{R}$ *let* $M_\theta\phi : \mathbb{N}\cup\{0\} \to \mathbb{R}$ *denote the function defined by* $(M_\theta\phi)(n) = \theta^n\phi(n)$ *and let* $\delta_\theta : T(V) \to T(V)$ *be the linear operator defined by* (2.7). *Then*:

(1) *For every* $a, b \in T(V)$, *we have the identity*

$$\langle\delta_\theta a, b\rangle_\phi = \langle a, \delta_\theta b\rangle_\phi = \langle a, b\rangle_{M_\theta\phi}, \tag{2.8}$$

which extends to $a, b \in T_{|M_\theta\phi|}((V))$. *The map* δ_θ *defines an isomorphism between the Hilbert spaces* $T_{M_{\theta^2}|\phi|}((V))$ *and* $T_{|\phi|}((V))$;

(2) *For* $|\theta| \le 1$ *and* $\phi > 0$ *we have* $T_\phi((V)) \subseteq T_{M_{\theta^2}\phi}((V))$ *and* $\delta_\theta : T_\phi((V)) \to T_\phi((V))$ *is a bounded self-adjoint linear operator with operator norm* $\|\delta_\theta\| \le 1$;

(3) *For* $|\theta| > 1$ *and* $\phi > 0$, δ_θ *is a linear operator* $\delta_\theta : D(\delta_\theta) \to T_\phi((V))$ *with domain* $T_{M_{\theta^2}\phi}((V)) \subseteq D(\delta_\theta) \subset T_\phi((V))$. *If furthermore* ϕ *satisfies Condition* 1, *then* $D(\delta_\theta)$ *is dense in* $T_\phi((V))$ *and* δ_θ *is self-adjoint.*

PROOF. For item 1, the identity (2.8) follows from (2.5). The extension to the completion follows from the fact that $|\langle \delta_\theta a, b\rangle_\phi| \le \|a\|_{|M_\theta \phi|} \|b\|_{|M_\theta\phi|}$. That δ_θ is an isometry between the pre-Hilbert spaces $(T(V)/N, \langle\cdot,\cdot\rangle_{M_{\theta^2}|\phi|})$ and $(T(V)/N, \langle\cdot,\cdot\rangle_{|\phi|})$ follows from (2.8) and the identity $\delta_\theta^2 = \delta_{\theta^2}$:

$$\langle \delta_\theta a, \delta_\theta b\rangle_{|\phi|} = \langle a, \delta_\theta^2 b\rangle_{|\phi|} = \langle a, b\rangle_{M_{\theta^2}|\phi|},$$

which extends to the completion $T_{M_{\theta^2}|\phi|}((V))$. Surjectivity follows from the fact that $\delta_\theta(T(V)) = T(V)$ for any nonzero θ. For item 2, it is readily seen that $\|a\|_{M_{\theta^2}\phi} \le \|a\|_\phi$ when $|\theta| \le 1$ for all $a \in T(V)$, and hence that $T_\phi((V)) \subseteq T_{M_{\theta^2}\phi}((V))$. By item 1, we then have $\|\delta_\theta a\|_\phi \le \|a\|_\phi$, which then extends to $T_\phi((V))$. Self-adjointness follows from the identity

$$\langle \delta_\theta a, b\rangle_\phi = \sum_{k=0}^\infty \theta^k \phi(k) \langle a_k, b_k\rangle_k = \langle a, \delta_\theta b\rangle_\phi \quad \text{for all } a, b \in T_\phi((V)). \tag{2.9}$$

Finally, for item 3 we observe that $T_{M_{\theta^2}\phi}((V))$ is a linear subspace of $T_\phi((V))$ and then that $\delta_\theta(T_{M_{\theta^2}\phi}((V))) \subseteq T_\phi((V))$ using item 1. If ϕ satisfies Condition 1, then the domain of δ_θ contains the linear span of the set of signatures $\mathcal{S}$ (recall (2.3)), which is dense in $T_\phi((V))$. Self-adjointness is again a consequence of (2.9). □

As an immediate corollary, we obtain the following result, which we shall use repeatedly.

COROLLARY 2.11. *Suppose* $\theta \in \mathbb{R}$ *and let* $\phi : \mathbb{N} \cup \{0\} \to \mathbb{R}$ *be such that* $|\phi|$ *satisfies Condition* 1 *then*

$$K^{\gamma,\sigma}_{M_\theta\phi}(s,t) = K^{\theta\gamma,\sigma}_\phi(s,t) = K^{\gamma,\theta\sigma}_\phi(s,t)$$

for every $(s,t) \in [a,b] \times [a,b]$, *where* $\theta\gamma$ *and* $\theta\sigma$ *denote the paths obtained by the pointwise multiplication of* θ *with* γ *and* σ, *respectively. In particular if* $\phi \equiv 1$, *then* $K^{\gamma,\sigma}_{\theta\phi} =: K^{\gamma,\sigma}_\theta$ *satisfies*

$$K^{\gamma,\sigma}_\theta(s,t) = 1 + \theta \int_a^s \int_a^t K^{\gamma,\sigma}_\theta(s,t) \langle d\gamma_u, d\sigma_v\rangle.$$

PROOF. We use the fact that $\delta_\theta S(\gamma)_{s,t} = S(\theta\gamma)_{s,t}$ and the previous lemma to observe that

$$K^{\gamma,\sigma}_{\theta\phi}(s,t) = \langle S(\gamma)_{a,s}, S(\sigma)_{a,t}\rangle_{\theta\phi} = \langle \delta_\theta S(\gamma)_{a,s}, S(\sigma)_{a,t}\rangle_\phi = K^{\theta\gamma,\sigma}_\phi(s,t).$$

The fact that $K^{\theta\gamma,\sigma}_\phi(s,t) = K^{\gamma,\theta\sigma}_\phi(s,t)$ follows from the self-adjointness of δ_θ. □

3. Representing functions using weighted signature kernels. We consider the way in which the signature kernel, and its weighted generalisations can be used to represent functions on paths space, and we explore how this is influenced by different choices of weight functions. To do so, we will work on the space of unparameterised paths, which we denote by $\mathcal{C}_1$. We recall that this is the set of equivalence classes of continuous, bounded variation paths, which are defined over the fixed interval $[a,b]$, in which two paths γ and σ are equivalent if their signatures agree. One of the main results of [14] is that this notion of equivalence coincides with the path-level notion of tree-like equivalence. On the space $\mathcal{C}_1$, the signature becomes a well-defined and injective map from $\mathcal{C}_1$ into $T((V))$; as in the previous section, we denote its range by $\mathcal{S}$. Where possible in this section, we suppress reference to $[a, b]$ and write $S(\gamma)$ instead of $S(\gamma)_{a,b}$.

Let E be a linear subspace of $T((V))$, which contains the range of the signature map $\mathcal{S}$ and on which there is defined an inner product $\langle\cdot,\cdot\rangle_E$. Then we recall [27] that there exists a unique reproducing kernel Hilbert space (RKHS) $(\mathcal{H}_E, \langle\cdot,\cdot\rangle_{\mathcal{H}_E})$ associated to E such that:

(1) If $k_{[\gamma]}$ is defined by

$$k_{[\gamma]} : [\sigma] \mapsto \langle S(\gamma), S(\sigma)\rangle_E \in \mathbb{R}^{\mathcal{C}_1},$$

then the linear span of $\{k_{[\gamma]} : [\gamma] \in \mathcal{C}_1\} \subset \mathbb{R}^{\mathcal{C}_1}$ is a dense subset of $\mathcal{H}_E$.

(2) $\langle k_{[\gamma]}, k_{[\sigma]}\rangle_{\mathcal{H}} = \langle S(\gamma), S(\sigma)\rangle_E$ for every $[\sigma]$ and $[\gamma]$ in $\mathcal{C}_1$.

This is the RKHS associated to the kernel $k_E : \mathcal{C}_1 \times \mathcal{C}_1 \to \mathbb{R}$ given by

$$k_E\big([\gamma], [\sigma]\big) = \langle S(\gamma), S(\sigma)\rangle_E,$$

the associated feature map being $[\gamma] \mapsto k([\gamma], \cdot) = k_{[\gamma]}$. To simplify the notation, we will omit reference to equivalence classes and write k_γ and $k(\gamma, \sigma)$ instead of $k_{[\gamma]}$ and $k([\gamma], [\sigma])$.

In the setting of weighted signature kernels, we are interested in this construction when $E = T_\phi((V))$ for weight functions ϕ, which satisfy Condition 1. We write the induced RKHS as $\mathcal{H}_\phi$ and then, because $T_\phi((V))$ contains the linear span of $\mathcal{S}$ as a dense subspace, we have that

$$\mathcal{H}_\phi = \big\{k_h : h \in T_\phi((V))\big\} \quad \text{with} \quad k_h : \gamma \mapsto \langle h, S(\gamma)\rangle_\phi,$$

so that $h \mapsto k_h$ is the usual isomorphism between $T_\phi((V))$ and $T_\phi((V))^*$ and $\mathcal{H}_\phi = T_\phi((V))^*$.

We see from this discussion that changing the choice of weight function leads to a possibly different RKHS. Under quite general conditions, however, they will all share with the original signature kernel the property of being universal.

DEFINITION 3.1. Let $\mathcal{C}_1$ be equipped with a topology. Let $k : \mathcal{C}_1 \times \mathcal{C}_1 \to \mathbb{R}$ be a continuous, symmetric, positive definite kernel on $\mathcal{C}_1$. Then we say that k is universal if for every compact subset $\mathcal{K} \subset \mathcal{C}_1$ the linear span of the set $\{k_\gamma : \gamma \in \mathcal{K}\}$ is dense in $C(\mathcal{K})$ with respect to the topology of uniform convergence.

REMARK 3.2. This notion is sometimes called cc-universality; see [27]. From the same reference, we know that k_ϕ is universal if and only if for any compact subset $\mathcal{K} \subset \mathcal{C}_1$, the set $\{k_h|_{\mathcal{K}} : h \in T_\phi((V))\}$ is dense in $C(\mathcal{K})$ in the uniform topology, where $k_h|_{\mathcal{K}}$ denotes the restriction of k_h to $\mathcal{K}$.

The argument for showing universality of a general signature kernels is an uncomplicated variation of the one for the original signature kernel; cf. [9]. We include a proof for completeness.

PROPOSITION 3.3. *Let $\phi : \mathbb{N} \cup \{0\} \to \mathbb{R}_{>0}$ be a weight function. Let $\mathcal{H}_\phi$ denote the reproducing kernel Hilbert space associated ϕ with the corresponding kernel denoted by k_ϕ. Suppose that $\mathcal{C}_1$ is equipped with a topology such that*:

(1) *the weight function ϕ satisfies Condition 1, and*
(2) *the signature map $S : \mathcal{C}_1 \to T_\phi((V))$ is continuous.*

Then k_ϕ is universal in the sense of Definition 3.1.

PROOF. The continuity and symmetry of $k_\phi : \mathcal{C}_1 \times \mathcal{C}_1 \to \mathbb{R}$ follows from the second assumption and $k_\phi(\gamma, \sigma) = \langle S(\gamma), S(\sigma)\rangle_\phi$. Taking advantage of Remark 3.2, it suffices to prove that $\{k_h|_\mathcal{K} : h \in T_\phi((V))\}$ is dense in $C(\mathcal{K})$. To do so, we note that for any h and g in $T(V)$ and γ in $\mathcal{K}$ it holds that

$$\langle h, S(\gamma)\rangle_\phi \langle g, S(\gamma)\rangle_\phi = \langle h \sqcup\!\sqcup_\phi g, S(\gamma)\rangle_\phi,$$

where $\sqcup\!\sqcup_\phi : T(V) \times T(V) \to T(V)$ is the bilinear map detemined by

$$h \sqcup\!\sqcup_\phi g = \frac{\phi(n)\phi(m)}{\phi(n+m)} h \sqcup\!\sqcup g \quad \text{for } h \in V^{\otimes n}, g \in V^{\otimes m}.$$

Note that when $\phi \equiv 1$ this is just the usual shuffle product. From this, it follows that $\{k_h|_\mathcal{K} : h \in T_\phi((V))\}$ contains the algebra $\mathcal{A} = \{k_h|_\mathcal{K} : h \in T(V)\}$. We can then conclude by using the Stone–Weierstrass theorem, since $\mathcal{A}$ can be seen to contain the constant function $k_{\mathbf{1}}|_\mathcal{K}$ and also to separate points in $\mathcal{K}$. □

REMARK 3.4. There are different ways to choose a topology on $\mathcal{C}_1$; see, for example, [5].

Despite all the kernels in this class having RKHSs, which are dense in $C(\mathcal{K})$, it is important to appreciate that the spaces $\mathcal{H}_\phi$ themselves will be different according to the choice of ϕ. A consideration in the selection of an appropriate kernel should take into account not only its universality, but also the efficiency with which a given RKHS can represent functions of interest. For general signature kernels, any weight function $\phi : \mathbb{N} \cup \{0\} \to \mathbb{R}_{>0}$, which satisfies $\phi \geq 1$ and Condition 1, will have an RKHS that contains the RKHS $\mathcal{H}$ of the original signature kernel. In principle therefore, functions which can be represented by a single element of $\mathcal{H}_\phi$ may have a much more complex (and only approximate) representation using elements of $\mathcal{H}$. The following example of a function on $\mathcal{C}_1$, which we return to later in a different setting, illustrates this point for a specific choice of an inner product.

EXAMPLE 3.5 (Hyperbolic development map). Let $V = \mathbb{R}^d$ with the Euclidean inner product. Suppose that $F : \mathbb{R}^d \to M_{d+1}(\mathbb{R})$ is the linear map into the space of $d+1$ by $d+1$ real matrices, which is defined by

$$F(x) = \begin{pmatrix} 0 & x \\ x^T & 0 \end{pmatrix}.$$

There exists a unique solution $y = (y^{i,j})_{i,j=1,\dots,d+1}$ in $M_{d+1}(\mathbb{R})$ to the differential equation

$$dy_t = F(d\gamma_t) \cdot y_t \quad \text{started at } y_0 = I_{d+1} \text{ with } t \in [0,1],$$

where $\cdot$ denotes matrix multiplication and I_{d+1} is the identity matrix in $M_{d+1}(\mathbb{R})$. The real-valued function defined by

$$k : \gamma \mapsto y_1^{d+1,d+1}$$

is invariant on tree-like equivalence classes and it therefore induces a function $\kappa : \mathcal{C}_1 \to \mathbb{R}$ by $\kappa([\gamma]) = k(\gamma)$. This function has the form

$$\kappa([\gamma]) = \langle \mathbb{E}[S(\circ B)_{0,1}], S(\gamma)\rangle_\phi \quad \text{with} \quad \phi(k) := 2^{k/2}\left(\frac{k}{2}\right)! \quad \text{for } k \in \mathbb{N} \cup \{0\},$$

where $S(\circ B)_{0,1}$ denotes the Stratonovich signature of d-dimensional standard Brownian motion. Using the Fawcett–Victoir formula [12], we have that

$$A = \mathbb{E}[S(\circ B)_{0,1}] = \exp\left(\frac{1}{2}\sum_{i=1}^{d} e_i^2\right),$$

which is easily verified to be in $T_\phi((V))$, and hence κ is in $\mathcal{H}_\phi$. On the other hand, κ is not in the RKHS $\mathcal{H}$ because there exists no $\tilde{A}$ in $T_{\phi\equiv1}((V))$, which allows κ to be realised as $\kappa([\gamma]) = \langle \tilde{A}, S(\gamma)\rangle$. Indeed, if such $\tilde{A} = \sum_{k\geq 0} \tilde{A}_k$ were to exist it would need to satisfy

$$\tilde{A}_{2k} = 2^k k! \pi_{2k} \exp\left(\frac{1}{2}\sum_{i=1}^d e_i^2\right) = \left(\sum_{i=1}^d e_i^2\right)^k,$$

where $\pi_n : T((V)) \mapsto V^{\otimes n}$ is the canonical projection. If this were to hold, then we would have $\sum_{k=0}^\infty \|\tilde{A}_{2k}\|^2 = \infty$, and clearly no such $\tilde{A}$ is in $T_{\phi\equiv1}((V))$.

3.1. *MMD distances.* Given a class of functions $\mathcal{F}$ and a topology on $\mathcal{C}_1$, one can attempt to define a notion of distance on probability measures on $\mathcal{C}_1$ by setting

$$d(\mu, \nu) = \sup_{f\in\mathcal{F}} \int_{\mathcal{C}_1} f \, d(\mu - \nu).$$

If $\mathcal{F}$ is the unit ball of a reproducing kernel Hilbert space H of a kernel with feature map Φ, then it is well known that these maximum mean discrepancy (MMD) distances have the equivalent form

$$d(\mu, \nu) = \|\mathbb{E}_{\Gamma\sim\mu}[\Phi(\Gamma)] - \mathbb{E}_{\Sigma\sim\nu}[\Phi(\Sigma)]\|_H,$$

which is more amenable to computation. For the ϕ-signature kernel, this means that $d_\phi(\mu, \nu)$ will equal

$$\begin{aligned} &\langle \mathbb{E}_{\Gamma\sim\mu}[S(\Gamma)], \mathbb{E}_{\Gamma\sim\mu}[S(\Gamma)]\rangle_\phi - 2\langle \mathbb{E}_{\Gamma\sim\mu}[S(\Gamma)], \mathbb{E}_{\Sigma\sim\nu}[S(\Sigma)]\rangle_\phi \\ &\quad + \langle \mathbb{E}_{\Sigma\sim\nu}[S(\Gamma)], \mathbb{E}_{\Sigma\sim\nu}[S(\Sigma)]\rangle_\phi. \end{aligned} \tag{3.1}$$

This approach, and variants of it, have been used in the kernel learning to propose statistics for goodness-of-fit tests; see, for example, [10]. The goal in these tasks is to use d_ϕ to design the rejection regions for hypothesis tests to determine whether an observed sample is drawn from a known target distribution μ. The model we focus on, motivated by a problem in radio astronomy, is where μ is determined by the Wiener measure $\mathcal{W}$ on path space. In this case, the result of [8] ensures that d_ϕ can distinguish two distributions in the sense that $d_\phi(\mathcal{W}, \nu) = 0$ if and only if $\mathcal{W} = \nu$. In practice, $d_\phi(\mathcal{W}, \nu)$ will be estimated from a sample $\{\gamma_1, \dots, \gamma_n\}$ drawn from the unknown distribution ν and for this approach to be viable, we need to be able to compute the expression in (3.1). This involves especially handling the term

$$\langle \mathbb{E}_{\Gamma\sim\mathcal{W}}[S(\Gamma)], S(\gamma)\rangle_\phi.$$

In Section 6, we will show Example 3.5 can be extended to derive a closed-form formula for this expression across a selection of weighted signature kernels.

4. Representing general signature kernels. In the previous two sections, we introduced the definition of the ϕ-signature kernel of continuous paths γ and σ to be the function $K_\phi^{\gamma,\sigma}(s,t)$. This amounts to reweighting the terms in the signature to give more or less emphasis to high-order terms compared to the original signature kernel, that is, $\langle\cdot,\cdot\rangle_\phi$ for $\phi \equiv 1$. In the present section, we will build an approach to representing ϕ-signature kernels in such a way that allows for efficient computation. The same idea is presented in multiple guises and then specialised within each case to yield particular examples. Before we present this method for ϕ-signature kernels, we consider the error estimates, which arise using a naive truncation-based approach.

4.1. *Truncated signature kernels.* In this subsection, we give an error estimate of the truncated ϕ-signature kernel and the full ϕ-signature kernel of two continuous bounded variation paths.

Let the truncated signature kernel be denoted

$$K_{\phi}^{(N)}(s,t) := \sum_{k=0}^{N} \phi(k)\langle S(\gamma)_{a,s}^{k}, S(\sigma)_{a,t}^{k}\rangle_{k} = \sum_{k=0}^{N} \phi(k) \sum_{|I|=k} S^{I}(\gamma)_{a,s} S^{I}(\sigma)_{a,t}. \tag{4.1}$$

We have the following proposition.

PROPOSITION 4.1. *Let $\gamma, \sigma : [a,b] \to V$ be two continuous paths of bounded variation. Assume that the function $\phi : \mathbb{N} \cup \{0\} \to \mathbb{R}$ is such that $|\phi|$ satisfies Condition 1, then the truncated signature kernel $K_{\phi}^{(N)}(s,t)$ converges to the ϕ-signature kernel $K_{\phi}^{\gamma,\sigma}(s,t)$ when N goes to infinity, and the error bound is*

$$\left|K_{\phi}^{\gamma,\sigma}(s,t) - K_{\phi}^{(N)}(s,t)\right| \le \sum_{k=N+1}^{\infty} |\phi(k)|\big(L_s(\gamma)L_t(\sigma)\big)^{k}(k!)^{-2}, \tag{4.2}$$

where $L_s(\gamma)$ is the length of the path segment $\gamma|_{[a,s]}$.

PROOF. By the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
\left|K_{\phi}^{\gamma,\sigma}(s,t) - K_{\phi}^{(N)}(s,t)\right| &\le \sum_{k=N+1}^{\infty} |\phi(k)|\left|\langle S(\gamma)_{a,s}^{k}, S(\sigma)_{a,t}^{k}\rangle_{k}\right| \\
&\le \sum_{k=N+1}^{\infty} |\phi(k)|\left\|S(\gamma)_{a,s}^{k}\right\|_{k}\left\|S(\sigma)_{a,t}^{k}\right\|_{k} \\
&= \sum_{k=N+1}^{\infty} |\phi(k)|\frac{(L_s(\gamma)L_t(\sigma))^{k}}{(k!)^2}.
\end{aligned}$$

Since $|\phi|$ satisfies Condition 1, the error goes to 0 as $N \to \infty$. $\square$

We analyse two concrete examples that we will revisit later using other methods.

- The first example takes ϕ to be

$$\phi(k) := \left(\frac{k}{2}\right)! := \Gamma\left(\frac{k}{2}+1\right), \tag{4.3}$$

  where Γ denotes the Gamma fucntion. This example plays a role in Section 6 when we consider the expected signature of Brownian motion.
- The second example is

$$\phi(k) = \frac{\Gamma(m+1)\Gamma(k+1)}{\Gamma(k+m+1)}, \tag{4.4}$$

  where $m \in \mathbb{R}_{+}$. The case when $m = 0$, $\phi(k) \equiv 1$ corresponds to the original signature kernel, while $m = 1$ gives $\phi(k) = \frac{1}{k+1}$, which are the sequence of moments of a random variable that is uniformly distributed on [0, 1].

The following corollary specialises the previously obtained error estimate to these cases.

COROLLARY 4.2. *Let $\gamma, \sigma : [a, b] \to V$ be two continuous paths of bounded variation. Denote the length of the path segment $\gamma|_{[a,s]}$ as $L_s(\gamma)$.*

(1) *The ϕ-signature kernel under $\phi(k) = (\frac{k}{2})!$ is well-defined and there is a constant C such that*

$$
(4.5) \qquad \left|K_\phi^{\gamma,\sigma}(s,t) - K_\phi^{(N)}(s,t)\right| \le C\left(\frac{e}{2N+2}\right)^{N+1/2} e_{N+1}\big(L_s(\gamma)L_t(\sigma)\big),
$$

where $e_{N+1}(x) := \sum_{k=N+1}^{\infty} \frac{x^k}{k!}$.

(2) *The ϕ-signature kernel under $\phi(k) = \frac{\Gamma(m+1)\Gamma(k+1)}{\Gamma(k+m+1)}$ is well-defined and the error bound is*

$$
(4.6) \qquad \left|K_\phi^{\gamma,\sigma}(s,t) - K_\phi^{(N)}(s,t)\right| \le \frac{\Gamma(m+1)}{(L_s(\gamma)L_t(\sigma))^{\frac{m}{2}}} I_m^{(N+1)}\big(2\sqrt{L_s(\gamma)L_t(\sigma)}\big)
$$

in which $I_m^{(N+1)}(z) := (\frac{z}{2})^m \sum_{k=N+1}^{\infty} \frac{(\frac{1}{4}z^2)^k}{\Gamma(k+m+1)\Gamma(k+1)}$ is the tail of the series defining the modified Bessel function $I_m(z)$ of the first kind of order m.

PROOF. It is easy to see that these two functions ϕ satisfy Condition 1, which makes sure that the ϕ-signature kernels are well-defined. For the error bound (4.5), by the Stirling's approximation, there exist two constants C_1, C_2 such that

$$
C_1 x^{x+\frac{1}{2}} e^{-x} \le x! \le C_2 x^{x+\frac{1}{2}} e^{-x} \quad \forall x > 0.
$$

Then we have

$$
\frac{(\frac{k}{2})!}{k!} \le \frac{C_2}{\sqrt{2}C_1}\left(\frac{e}{2k}\right)^{\frac{k}{2}}
$$

and the sequence on the right-hand side is decreasing. Let $C = \frac{C_2}{\sqrt{2}C_1}$ and combine Proposition 4.1, it is easy to show the error bound (4.5).

For the error bound (4.6), since the modified Bessel function $I_m(2\sqrt{L_s(\gamma)L_t(\sigma)})$ of the first kind of order m is defined by the series

$$
I_m\big(2\sqrt{L_s(\gamma)L_t(\sigma)}\big) = \big(L_s(\gamma)L_t(\sigma)\big)^{\frac{m}{2}} \sum_{k=0}^{\infty} \frac{(L_s(\gamma)L_t(\sigma))^k}{\Gamma(k+m+1)\Gamma(k+1)},
$$

the error bound follows from Proposition 4.1. □

4.2. *General signature kernels by randomisation.* We now show how ϕ-signature kernels can be represented, under suitable integrability conditions, as the average of rescaled PDE solutions whenever the sequence $\{\phi(k) : k = 0, 1, \dots\}$ coincides with the sequence of moments of a random variable. This representation consolidates the connection between the original and the ϕ-signature kernels in these cases. The connection is captured in the following result.

PROPOSITION 4.3. *Suppose π is a random variable with finite moments of all orders and let the functions*

$$
(4.7) \qquad \phi(k) = \mathbb{E}[\pi^k] \quad \text{and} \quad \psi(k) = \mathbb{E}[|\pi|^k] \quad \forall k \ge 0.
$$

We assume that ψ satisfies Condition 1. Then the ϕ-signature kernel $K_\phi^{\gamma,\sigma}(s,t)$ of continuous bounded variation paths γ and σ is well-defined and

$$
(4.8) \qquad K_\phi^{\gamma,\sigma}(s,t) = \mathbb{E}_\pi\big[K^{\pi\gamma,\sigma}(s,t)\big] = \mathbb{E}_\pi\big[K^{\gamma,\pi\sigma}(s,t)\big].
$$

PROOF. Since $|\phi|$ satisfies Condition 1, which follows from the condition of ψ, the ϕ-signature kernel $K_\phi^{\gamma,\sigma}(s,t)$ is well-defined. Furthermore, ψ satisfies Condition 1, by Fubini's theorem, we have

$$\begin{aligned} K_\phi^{\gamma,\sigma}(s,t) &= \sum_{k=0}^{\infty} \mathbb{E}[\pi^k] \langle S(\gamma)_{a,s}^k, S(\sigma)_{a,t}^k \rangle_k \\ &= \mathbb{E}\Bigg[\sum_{k=0}^{\infty} \pi^k \langle S(\gamma)_{a,s}^k, S(\sigma)_{a,t}^k \rangle_k\Bigg] \\ &= \mathbb{E}\Bigg[\sum_{k=0}^{\infty} \langle S(\pi\gamma)_{a,s}^k, S(\sigma)_{a,t}^k \rangle_k\Bigg] \\ &= \mathbb{E}\big[K^{\pi\gamma,\sigma}(s,t)\big]. \end{aligned}$$

We conclude the proof. □

REMARK 4.4. If the random variable π has a known probability density function, the expectation in equation (4.8) can be calculated by numerical methods such as the Monte Carlo method or Gaussian quadrature procedure.

The corollary below gives two specialisations of this result to the cases described earlier.

COROLLARY 4.5. *Let $\gamma, \sigma : [a,b] \to V$ be two continuous paths of bounded variation.*

(1) *The ϕ-signature kernel under $\phi(k) = (\frac{k}{2})!$ satisfies*

$$K_\phi^{\gamma,\sigma}(s,t) = \mathbb{E}_\pi\big[K^{\pi^{1/2}\gamma,\sigma}(s,t)\big] = \mathbb{E}_\pi\big[K^{\gamma,\pi^{1/2}\sigma}(s,t)\big], \tag{4.9}$$

where $\pi \sim \mathrm{Exp}(1)$ is an exponentially distributed random variable with intensity 1.

(2) *The ϕ-signature kernel $K_\phi^{\gamma,\sigma}(s,t)$ under $\phi(k) = \frac{\Gamma(m+1)\Gamma(k+1)}{\Gamma(k+m+1)}$ satisfies equation* (4.8) *where $\pi \sim B(1,m)$ is a Beta-distributed random variable.*

PROOF. For (1), we need to show that ϕ is all the moments of the random variable $\pi^{1/2}$. Since $\pi \sim \mathrm{Exp}(1)$, we have

$$\mathbb{E}[\pi^{k/2}] = \int_0^\infty x^{k/2} e^{-x}\,dx = \Gamma\left(\frac{k}{2}+1\right) = \phi(k).$$

Equation (4.9) then follows from Theorem 4.3. For (2), since the random variable π is Beta distributed, that is, $\pi \sim \mathrm{Beta}(1,m)$, then the moments of π are

$$\mathbb{E}[\pi^k] = \frac{B(k+1,m)}{B(1,m)} = \frac{\Gamma(k+1)\Gamma(m+1)}{\Gamma(k+m+1)} = \phi(k).$$

We conclude the proof. □

The motivation for the representation (4.8) is that we can design efficient and accurate computational methods to compute the ϕ-signature kernels. We will give details on the Gaussian quadrature methods for the ϕ-signature kernel in Section 5 below.

4.3. *General signature kernels by Fourier series.* We now extend the earlier discussion so that $\phi : \mathbb{Z} \to \mathbb{C}$ is a complex-valued function. We consider the blinear form defined by the two-sided summation

$$\langle a, b\rangle_\phi := B_\phi(a, b) := \sum_{k=-\infty}^{\infty} \phi(k)\langle a_{|k|}, b_{|k|}\rangle_{|k|},$$

and the corresponding function

$$K_\phi^{\gamma,\sigma}(s, t) := \langle S(\gamma)_{a,s}, S(\sigma)_{a,t}\rangle_\phi.$$

If the coefficients are the Fourier coefficients of some known periodic function f then the idea of the previous proposition can be applied to again derive a representation of $K_\phi^{\gamma,\sigma}$. The following result describes the needed conditions.

PROPOSITION 4.6. *Suppose that γ and σ are continuous paths of bounded* 1*-variation. Let $\phi : \mathbb{Z} \to \mathbb{C}$ be as above, and write $\phi_k := \phi(k)$. Assume that $\{\phi_k : k \in \mathbb{N}\}$ are the Fourier coefficients of some bounded integrable function $f : (-\pi, \pi) \to \mathbb{C}$, that is,*

$$f = \sum_{k=-\infty}^{\infty} \phi_k e^{ikx}.$$

Then for all $(s, t) \in [a, b] \times [a, b]$, we have

$$K_\phi^{\gamma,\sigma}(s, t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \bar{K}_x^{\gamma,\sigma}(s, t) f(x)\, dx - \phi_0, \tag{4.10}$$

where

$$\bar{K}_x^{\gamma,\sigma}(s, t) := K^{\exp(-ix)\gamma,\sigma}(s, t) + K^{\exp(ix)\gamma,\sigma}(s, t).$$

PROOF. Fixing (s, t), we have for every $x \in (-\pi, \pi)$ that

$$K^{\exp(\pm ix)\gamma,\sigma}(s, t) = \sum_{k=0}^{\infty} e^{\pm ikx}(x)\langle S(\gamma)_{a,s}^k, S(\sigma)_{a,t}^k\rangle_k =: \sum_{k=0}^{\infty} e^{\pm ikx}(x) c_k.$$

The basic estimate $|c_k| \le L_\gamma^k L_\sigma^k/(k!)^2$, where L_γ is the length of the path γ, ensures that $\sum_{k=0}^N c_k e^{\pm ikx}(\cdot) f(\cdot)$ converges uniformly to the series $\sum_{k=0}^\infty c_k e^{\pm ikx}(\cdot) f(\cdot)$, and hence

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} K^{\exp(\pm ix)\gamma,\sigma}(s, t) f(x)\, dx = \frac{1}{2\pi}\sum_{k=0}^{\infty} c_k \int_{-\pi}^{\pi} e^{\pm ikx}(x) f(x)\, dx = \sum_{k=0}^{\infty} c_k \phi_{\mp k}.$$

It follows that

$$\begin{aligned}&\frac{1}{2\pi}\int_{-\pi}^{\pi}\big[K^{\exp(-ix)\gamma,\sigma}(s, t) + K^{\exp(ix)\gamma,\sigma}(s, t)\big] f(x)\, dx\\ &\quad = \sum_{k=-\infty}^{\infty} c_{|k|}\phi_k + c_0\phi_0 = K_\phi^{\gamma,\sigma}(s, t) + \phi_0,\end{aligned}$$

as required. □

REMARK 4.7. Note that $\mathcal{R}K_x^{\gamma,\sigma}(s, t) := \operatorname{Re} K^{\exp(ix)\gamma,\sigma}(s, t)$ is given by

$$\mathcal{R}K_x^{\gamma,\sigma}(s, t) = \sum_{k=0}^{\infty} \cos kx \langle S(\gamma)_{a,s}^k, S(\sigma)_{a,t}^k\rangle_k.$$

Together with $\mathcal{I}K_x^{\gamma,\sigma}(s,t) := \operatorname{Im} K^{\exp(ix)\gamma,\sigma}(s,t)$, it solves the 2-dimensional PDE:

$$\frac{\partial^2}{\partial s \partial t}\begin{pmatrix}\mathcal{R}K_x^{\gamma,\sigma}(s,t)\\ \mathcal{I}K_x^{\gamma,\sigma}(s,t)\end{pmatrix} = \begin{pmatrix}\cos x & -\sin x\\ \sin x & \cos x\end{pmatrix}\begin{pmatrix}\mathcal{R}K_x^{\gamma,\sigma}(s,t)\\ \mathcal{I}K_x^{\gamma,\sigma}(s,t)\end{pmatrix}\langle \gamma_s', \sigma_t'\rangle.$$

COROLLARY 4.8. *Special cases of the above result include*:

(1) *If* $\phi_k = 0$ *for* $k < 0$, *then*

$$K_\phi^{\gamma,\sigma}(s,t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} K^{\exp(-ix)\gamma,\sigma}(s,t) f(x)\, dx.$$

(2) (*Real Fourier series*) *Suppose*

$$f = a_0 + \sum_{k=1}^{\infty} a_k c_k + \sum_{k=1}^{\infty} b_k s_k \quad \textit{where } c_k(\cdot) := \cos(k\cdot), s_k(\cdot) := \sin(k\cdot)$$

with $\{a_k\}$ *and* $\{b_k\}$ *real sequences. If* $\phi(k) = a_k$ *so that*

$$\langle p, q\rangle_\phi := \sum_{k=0}^{\infty} a_k \langle p_k, q_k\rangle_k, \tag{4.11}$$

then

$$K_\phi^{\gamma,\sigma}(s,t) = \frac{1}{\pi}\int_{-\pi}^{\pi} \mathcal{R}K_x^{\gamma,\sigma}(s,t) f(x)\, dx - a_0.$$

In using this result, the function f should be chosen that the integral can easily approximated numerically.

EXAMPLE 4.9. The following simple examples illustrate the scope of these ideas:

(1) The function $f(x) = x^2$ has the Fourier series $f = \sum_{k=0}^{\infty} \phi_k c_k$ on $[-\pi, \pi]$ where

$$\phi_k = \frac{4(-1)^k}{k^2}, \qquad \phi_0 = \frac{\pi^2}{3},$$

and we obtain the identity

$$\frac{\pi^2}{3} + \sum_{k=1}^{\infty} \frac{4(-1)^k}{k^2}\langle S(\gamma)_{a,s}^k, S(\sigma)_{a,t}^k\rangle_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{R}K_x^{\gamma,\sigma}(s,t) x^2\, dx.$$

(2) The periodic function $f(x) = e^{\cos x}\cos(\sin x)$ has Fourier series

$$f(x) = \sum_{k=0}^{\infty} \frac{1}{k!}\cos(kx)$$

and so

$$1 + \sum_{k=0}^{\infty} \frac{1}{k!}\langle S(\gamma)_{a,s}^k, S(\sigma)_{a,t}^k\rangle_k = \frac{1}{\pi}\int_{-\pi}^{\pi} \mathcal{R}K_x^{\gamma,\sigma}(s,t) e^{\cos x}\cos(\sin x)\, dx.$$

(3) The Jacobi theta function is the 1-periodic function

$$\theta(z;\tau) = 1 + 2\sum_{k=1}^{\infty} e^{i\pi\tau k^2}\cos(2\pi k z),$$

hence if we define $f(x;u) := \theta(\frac{x}{2\pi}; \frac{iu}{\pi})$, then $f(\cdot;u) = 1 + \sum_{k=1}^{\infty} e^{-uk^2} c_k$ and

$$\sum_{k=0}^{\infty} e^{-uk^2}\langle S(\gamma)_{a,s}^k, S(\sigma)_{a,t}^k\rangle_k = \frac{1}{\pi}\int_{-\pi}^{\pi} \mathcal{R}K_x^{\gamma,\sigma}(s,t) f(x;u)\, dx - 1.$$

4.4. *General signature kernels by integral transforms.* The main idea of the previous subsection was to look for a function f with Fourier series $\sum_{k\in\mathbb{Z}}\phi(k)e^{ikx}$. If such a function can be found then, in principle, we can calculate the bilinear form B_ϕ evaluated at a pair of signatures. The difficulty with this approach is that such a function may not exist in some cases of interest, for example, $\phi(k)=k^{-1/2}$, $\phi(k)=k!$, etc. To simplify, we forego the two-sided summation and re-define

$$\langle a,b\rangle_\phi := B_\phi(a,b) := \sum_{k=0}^{\infty}\phi(k)\langle a_k,b_k\rangle_k,$$

where $\phi:\mathbb{R}\to\mathbb{C}$ is now defined on $\mathbb{R}$. To capture more generally some of the structure used above, we assume now that ϕ is the integral of a function $r:\mathbb{R}\times\mathbb{R}\to\mathbb{C}$ against a finite signed Borel measure μ on $\mathbb{R}$ such that

$$\phi(u)=\int_C = g(z)^{\alpha u}\mu(dz) \quad \text{for } \alpha\in\mathbb{R}. \tag{4.12}$$

The pairs (ϕ,μ) related in this way include well-known examples of integral transforms.

EXAMPLE 4.10. We will consider three principal examples:

(1) Fourier–Stieltjes transform: $C=\mathbb{R}$, $g(z)=e^{-2\pi iz}$, $\alpha=1$, that is, $\phi(u)=\hat{\mu}(u):=\int_{\mathbb{R}}e^{-2\pi iuz}\mu(dz)$;

(2) Laplace–Stieltjes transform: $C=(0,\infty)$, $g(z)=e^{-z}$, $\alpha=1$, that is, $\phi(u)=\tilde{\mu}(u):=\int_0^\infty e^{-uz}\mu(dz)$;

(3) Mellin–Stieltjes transform: $C=(0,\infty)$, $g(z)=z$, $\alpha=1$, that is, $\phi(u)=\mu_{\mathrm{Mel}}(u+1)=\int_0^\infty z^u\mu(dz)$, $\mathrm{Re}\,u>-1$.

In the general case, we can expect—under reasonable assumptions—that the integral representation can be used to justify the calculation

$$\begin{aligned}\langle S(\gamma)_{a,s},S(\sigma)_{a,t}\rangle_\phi &= \sum_{k=0}^{\infty}\int_C g(z)^{\alpha k}\mu(dz)\langle S(\gamma)^k_{a,s},S(\sigma)^k_{a,t}\rangle_k\\ &= \int_C\sum_{k=0}^{\infty}\langle S\big(g(z)^\alpha\gamma\big)^k_{a,s},S(\sigma)^k_{a,t}\rangle_k\mu(dz)\\ &= \int_C K^{g(z)^\alpha\gamma,\sigma}(s,t)\mu(dz),\end{aligned} \tag{4.13}$$

again allowing us to reduce the calculation of the bilinear form to a weighted integral over PDE solutions. On this occasion, integration is w.r.t. the measure μ and the rescaling is determined by the form of the kernel function r in the integral transform relating μ and ϕ.

THEOREM 4.11. *Let μ be a finite signed Borel measure μ on $\mathbb{R}$. Suppose that $\phi:\mathbb{R}\to\mathbb{C}$ is such that*

$$\phi(k)=\int_C g(z)^{\alpha k}\mu(dz)\in\mathbb{C} \quad \textit{for all } k\in\mathbb{N}\cup\{0\}$$

for a function $g:\mathbb{C}\to\mathbb{C}$. Let

$$\gamma:[a,b]\to V \quad \textit{and} \quad \sigma:[a,b]\to V$$

be continuous paths of bounded 1-variation with signatures $S(\gamma)$ and $S(\sigma)$, respectively. For every $(s,t)\in[a,b]\times[a,b]$ and $k\in\mathbb{N}\cup\{0\}$, define

$$a_k(s,t):=\langle S(\gamma)^k_{a,s},S(\sigma)^k_{a,t}\rangle_k.$$

Assume for every $(s,t)\in[a,b]\times[a,b]$ that:

(1) *the integral* $\int_C |g(z)^{\alpha k})||\mu(dz)| < \infty$, *and*
(2) *the series* $\sum_k a_k(s,t) \int_C |g(z)^{\alpha k})||\mu(dz)|$ *converges absolutely*,

then

$$\langle S(\gamma)_{a,s}, S(\sigma)_{a,t}\rangle_\phi = \int_C K^{g(z)^\alpha \gamma, \sigma}(s,t)\mu(dz). \tag{4.14}$$

REMARK 4.12. Sufficient for item 2 is that $\sum_k A^k (k!)^{-2} \int_C |h_k(z; s,t)||dz|$ converges for every $A > 0$.

PROOF. Assumptions 1 and 2 above ensure that Fubini's theorem can be applied to give

$$\sum_{k=0}^{\infty} a_k(s,t) \int_C g(z)^{\alpha k}\mu(dz) = \int_C \sum_{k=0}^{\infty} a_k(s,t) g(z)^{\alpha k}\mu(dz),$$

which can be seen to be the same as (4.14). $\square$

COROLLARY 4.13. *For each of the three integral transforms in Example* 4.10 *satisfying assumption* 1 *and* 2 *in the above theorem, we have* (4.14).

In a similar way we have the following results once again.

COROLLARY 4.14. *Let* π *be a complex-valued random variable with finite moments of all orders and*

$$\phi(k) = \mathbb{E}[\pi^k] \quad \text{and} \quad \psi(k) = \mathbb{E}[|\pi|^k] \quad \text{for all } k \in \mathbb{N} \cup \{0\}$$

such that ψ *satisfies Condition* 1. *Then*

$$K_\phi^{\gamma,\sigma}(s,t) = \mathbb{E}_\pi[K^{\pi\gamma,\sigma}(s,t)] = \mathbb{E}_\pi[K^{\gamma,\pi\sigma}(s,t)].$$

PROOF. Let F be the distribution function of π. Apply Theorem 4.11 with $\mu = dF$ and $r(u,z) = z^u$. $\square$

EXAMPLE 4.15. These examples illustrate these results:

(1) For any $\beta > -1$, the function $\phi(u) = \Gamma(u+\beta+1) = \int_0^\infty x^u x^\beta e^{-x}\, dx$ is the Mellin transform of $x^\beta e^{-x}$. Therefore, we have

$$\sum_{k=0}^{\infty} \Gamma(k+\beta+1)\langle S(\gamma)_{a,s}^k, S(\sigma)_{a,t}^k\rangle_k = \int_0^\infty K^{x\gamma,\sigma}(s,t) x^\beta e^{-x}\, dx.$$

(2) Suppose π is a random variable, the expectation can be computed in the following cases:
   (a) if π is uniformly distributed on $[0,1]$, then it equals

$$\sum_{k=0}^{\infty} \frac{1}{k+1}\langle S(\gamma)_{a,s}^k, S(\sigma)_{a,t}^k\rangle_k = \int_0^1 K^{x\gamma,\sigma}(s,t)\, dx;$$

   (b) if π has the Arcsine$(-1,1)$-distribution, that is, $F_\pi(x) = \frac{2}{\pi}\arcsin(\sqrt{\frac{1+x}{2}})$, then

$$\sum_{k=0}^{\infty} \prod_{r=0}^{2k-1} \frac{2r+1}{2r+2}\langle S(\gamma)_{a,s}^k, S(\sigma)_{a,t}^k\rangle_k = \frac{1}{\pi}\int_{-1}^1 \frac{K^{x\gamma,\sigma}(s,t)}{\sqrt{1-x^2}}\, dx; \tag{4.15}$$

   (c) if π has the Beta(α,β)-distribution, then

$$\sum_{k=0}^{\infty} \prod_{r=0}^{k-1} \frac{\alpha+\beta}{\alpha+\beta+r}\langle S(\gamma)_{a,s}^k, S(\sigma)_{a,t}^k\rangle_k = \frac{1}{B(\alpha,\beta)}\int_0^1 K^{x\gamma,\sigma}(s,t) x^{\alpha-1}(1-x)^{\beta-1}\, dx.$$

5. Computing general signature kernels. The usefulness of the formulae in the last section depends on one being able to numerically approximate integrals such as

$$\int_a^b f(x)w(x)\,dx,$$

where $[a,b] \subseteq \mathbb{R}$, $w \in L^1((a,b))$ is a weight function, which for the moment we assume to be positive. In the examples considered, the function f to be integrated will be a scaling of the signature kernel PDE, and typically we will have

$$f(x) = K^{x\gamma,\sigma}(s,t).$$

A classical approach to such approximations is to use a Gaussian quadrature rule; see, for example, [30].

For a general weight function w, suppose that $\mathcal{P} = \{p_n : n \in \mathbb{N} \cup \{0\}\}$ is a system of orthogonal polynomials w.r.t. the weight function w over (a,b); that is, $\deg(p_n) = n$ and $\langle p_n, p_m\rangle_w = \int_a^b p_m p_n w\,dx = 0$ for $n \neq m$. Then the quadrature points x_k, $k = 0, 1, \dots, n$ are the zeros of the polynomial p_{n+1}, the corresponding quadrature weights are

$$w_k := \int_a^b w(x) \prod_{i=0, i\neq k}^{n} \left(\frac{x - x_i}{x_k - x_i}\right)^2 dx$$

and the quadrature rule is the approximation

$$\int_a^b f(x)w(x)\,dx \approx \sum_{k=0}^{n} w_k f(x_k).$$

The approximation is exact if f is a polynomial with $\deg(f) \leq 2n+1$. If f is assumed to be C^{2n+2}, then the error in the quadrature rule can be approximated by the basic estimate [30],

$$\left|\int_a^b f(x)w(x)\,dx - \sum_{k=0}^{n} w_k f(x_k)\right| \leq \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\int_a^b w(x)\pi_{n+1}(x)^2\,dx, \tag{5.1}$$

where $\xi \in (a,b)$ and

$$\pi_{n+1}(x) = \prod_{i=0}^{n}(x - x_i)$$

is the monic poynomial obtained by dividing p_{n+1} by its leading coefficient. In view of the bound (5.1), it is useful to have estimates on the derivatives of the function $x \mapsto K^{x\gamma,\sigma}(s,t)$. To this end, we have the following.

LEMMA 5.1. *Define $f(x) := K^{x\gamma,\sigma}(s,t)$ for $x \in \mathbb{R}$. Then f is infinitely differentiable and, for every $k \in \mathbb{N}$, its kth derivative is given by*

$$f^{(k)}(x) = \sum_{l=0}^{\infty} x^l \frac{(l+k)!}{l!}\langle S(\gamma)_{a,s}^{l+k}, S(\sigma)_{a,t}^{l+k}\rangle_{l+k}. \tag{5.2}$$

In particular, we have the estimate

$$\left|f^{(k)}(x)\right| \leq \frac{L_s(\gamma)^{k/2}L_t(\sigma)^{k/2}}{|x|^{k/2}} I_k\bigl(2\sqrt{|x|L_s(\gamma)L_t(\sigma)}\bigr), \tag{5.3}$$

where $L_s(\gamma)$ is the length of the path segment $\gamma|_{[a,s]}$ and I_k is the modified Bessel function of the first kind of order k.

PROOF. Differentiablity is a simple argument on term-by-term differentiation of power series. Applying this argument k times results in the formula (5.2). The bound (5.3) can be obtained by the elementary estimate

$$\begin{aligned}\left|f^{(k)}(x)\right| &\leq \sum_{l=0}^{\infty} |x|^l \frac{(l+k)!}{l!} \frac{L_s(\gamma)^{l+k} L_t(\sigma)^{l+k}}{(l+k)!^2} \\ &= \frac{L_s(\gamma)^{k/2} L_t(\sigma)^{k/2}}{|x|^{k/2}} I_k\big(2\sqrt{|x| L_s(\gamma) L_t(\sigma)}\big).\end{aligned}$$

□

REMARK 5.2. For any $x \in \mathbb{R}$, $k \in \mathbb{N}$, it is easy to derive from (5.2) the crude estimate

$$\left|f^{(k)}(x)\right| \leq \frac{L_s(\gamma)^k L_t(\sigma)^k}{k!} \exp\big(|x| L_s(\gamma) L_t(\sigma)\big),$$

which could be refined, for example, by considering estimate on ratios of Bessel functions I_{k+1}/I_k.

Putting things together, we obtain the following estimate for the quadrature error.

PROPOSITION 5.3. *Let $\mathcal{P} = \{p_n : n \in \mathbb{N} \cup \{0\}\}$ be a system of orthogonal polynomials with respect to a continuous positive weight function $w \in L^1(a,b)$. For every n, the error in the associated quadrature formula satisfies the estimate*

$$\begin{aligned}&\left|\int_a^b K^{x\gamma,\sigma}(s,t) w(x)\,dx - \sum_{k=0}^{n} w_k K^{x_k\gamma,\sigma}(s,t)\right| \\ &\quad \leq \frac{L_s(\gamma)^{2n+2} L_t(\sigma)^{2n+2} \exp(|\xi| L_s(\gamma) L_t(\sigma))}{[(2n+2)!]^2} \int_a^b w(x) \pi_{n+1}(x)^2\,dx.\end{aligned}$$

EXAMPLE 5.4. Let $(a,b) = (-1,1)$, $w(x) = \frac{1}{\pi}\frac{1}{\sqrt{1-x^2}}$ as in the earlier example (4.15). Then $\mathcal{P}$ can be the family of Chebyshev polynomials of the first kind $p_n = T_n$ in which case (see [1])

$$\int_a^b w(x) \pi_{n+1}(x)^2\,dx = \frac{1}{2^{2n+1}}.$$

Therefore, if γ and σ have lengths at most L the degree $n+1$ quadrature rule results in an error at most

$$R_n = \frac{L^{4n+4} \exp(L^2)}{2^{2n+1}[(2n+2)!]^2}.$$

To give some idea of the number of points needed (and hence the number of PDEs solutions needed), if $L = 10$ then $R_{25} = e^{-8.6017}$, $R_{30} = e^{-50.492}$, whereas if $L = 100$ then $R_{1050} = e^{-49.497}$. The ratio

$$\frac{R_{n+1}}{R_n} = \frac{L^4}{4(2n+4)^2(2n+3)^2},$$

articulates the tradeoff between the length L and degree of the quadrature rule that is used.

EXAMPLE 5.5. The ϕ-signature kernel $K_\phi^{\gamma,\sigma}(s,t)$ for $\phi(k) = (\frac{k}{2})!$ is studied in Corollary 4.5. In this case, the random variable π is exponentially distributed, hence $\pi^{1/2}$ is Rayleigh distributed with density $w(x) = 2xe^{-x^2}$, $x > 0$. We have

$$K_\phi^{\gamma,\sigma}(s,t) = \mathbb{E}\big[K^{\pi^{1/2}\gamma,\sigma}(s,t)\big] = \int_0^\infty 2K^{x\gamma,\sigma}(s,t) x e^{-x^2}\,dx. \tag{5.4}$$

Let $f(x) = K^{x\gamma,\sigma}(s,t)$, then

$$K_\phi^{\gamma,\sigma}(s,t) = 2\int_0^\infty f(x)xe^{-x^2}\,dx,$$

which can be numerically calculated by the classical Gaussian quadrature formula (see, e.g., [26, 28]),

$$\int_0^\infty f(x)xe^{-x^2}\,dx \approx \sum_{k=0}^n w_k f(x_k).$$

The abscissae x_k, $k = 0, 1, \dots, n$ are the roots of a $(n+1)$-th degree polynomial $p_{n+1}(x)$ and w_k are the weights of quadrature. Explicit values are given in [26, 28].

5.1. *Quadrature versus truncation*: *A comparison.* We compare the relative accuracy of truncation-based methods with hybrid quadrature-PDE based methods. We do not discuss the complexity-based performance of the methods, although this question is important, too, and is the subject of ongoing research. As the quadrature error depends on the weight function defining the kernel, a separate analysis is needed for each kernel. We illustrate the comparison in the case where setting $\phi(k) = (k+1)^{-1}$; that is, we consider

$$\langle S(\gamma), S(\sigma)\rangle_\phi = \sum_{k=0}^\infty \frac{1}{k+1}\langle S_k(\gamma), S_k(\sigma)\rangle_k. \tag{5.5}$$

As the earlier estimate shows, the quadrature error depend on the lengths of the paths γ and σ. We therefore assume that the maximum of the lengths of the two paths γ and σ are bounded by L when, by the estimate in Proposition 4.1, the truncation error will be bounded by

$$\mathrm{TE}(m, L) := \sum_{k=m+1}^\infty \frac{L^{2k}}{(k+1)!k!}. \tag{5.6}$$

For this choice of weight function, the quadrature error is bounded by (see p. 888 of [1])

$$\mathrm{QE}(n, L) := \frac{[(n+1)!]^4}{(2n+3)[(2n+2)!]^3} L^{2n+2} I_{2n+2}(2L),$$

where we have made use of the estimates for the derivatives of the scaled signature kernel in Lemma 5.1. To obtain a comparison, we fix L and a level of accuracy ϵ, and then find the minimal m and n, respectively, for which $\mathrm{TE}(m, L) \le \epsilon$ and $\mathrm{QE}(n, L) \le \epsilon$. In the plot shown in Figure 1 below, we take $\epsilon = 10^{-6}$ and consider L in the range $[5, 120]$. The function $\mathrm{QE}(n, L)$ is evaluated numerically using the package `scipy.special`. To evaluate $\mathrm{TE}(m, L)$, the series (5.6) is estimated by truncation: simple calculus show that the summands are maximised when $k \approx L$ after which the series rapid converges, for example, for $r > e$,

$$\sum_{k=rL}^\infty \frac{L^{2k}}{(k+1)!k!} < \sum_{k=rL}^\infty \left(\frac{Le}{k}\right)^{2k} \le \left(\frac{e}{r}\right)^{2rL} \frac{r}{r-e},$$

so that $\mathrm{TE}(m, L)$ can be well approximated by retaining relatively few summands. The plot below is computed by setting $rL = 2000$ so that $16.667 \le r \le 400$, and the truncation of the series at rL gives a more than adequate estimate for $\mathrm{TE}(m, L)$ for the range of L considered and the choice of error tolerance ϵ.

We see qualitatively that the higher degrees of truncation/quadrature are needed as L increases, that this increase is approximately linear in L over the range considered, and that the number of terms needed using the truncation-based approach increases at a faster rate than

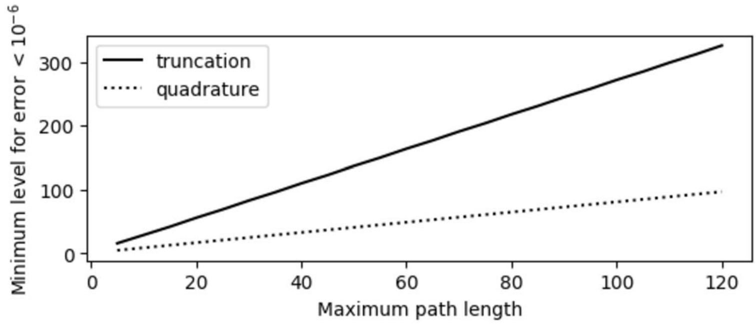

FIG. 1. *The minimum level of truncation and minimum-degree quadrature formula needed to achieve accuracy of* 10^{-6} *when computing the weighted signature kernel* (5.5) *for two paths, plotted against the maximum length of those paths.*

the degree of the quadrature formula. This observation remains stable over different choices of ϵ. The calculation of high-order iterated integrals is computationally nontrivial, and becomes increasingly expensive as the dimension d of the underlying state space increases. This provides evidence that the hydrid quadrature-signature PDE approach can be a useful alternative, even for moderate length paths ($L \approx 100$) such as those seen in applications (see below), where truncation-based schemes may be unfeasible.

5.2. *Weighted signature kernels for time-series classification.* To illustrate the use of weighted signature kernels, we apply them to the challenge of multivariate time series classification using the UEA data sets, which are available at https://timeseriesclassification.com/. We use the support vector classifier (SVC) as [23, 33], and compare accuracy-based performance under the same SVC settings (time-series preprocessing, hyperparameter selection, etc.) for the original signature kernel, the factorially-weighted signature kernel (see Example 5.5) and Beta-weighted signature kernel (here, we use $m = 1$ for the Beta weights). All signature kernels are computed using the PDE method of [23]. For the weighted signature kernels this is combined, in the case of the factorially-weighted signature kernel, with the use of a degree 16 quadrature rule ($n = 15$ in Example 5.5) with the weights and abscissae given on page 316, Table IIb, of [26]. For the Beta-weighted kernel, the PDE method is combined with the representation formula (4.8), the expectation being approximated using 21 equally-spaced points from the interval [0, 1]. Table 1 shows the performance of the SVC with the different kernels. We run the experiments both with and without augmenting the time series by adding an extra time coordinate. The numbers in bold typeface indicate in which experiments the test accuracy of the SVC using the factorially- or Beta-weighted signature kernel outperforms the original signature kernel.

6. Expected general signature kernels. We develop our earlier discussion to consider how ϕ-signature kernels can be combined with the notion of expected signatures to compare the laws of two stochastic processes. In the examples that we study, one of the measures will be Wiener's measure, which we denote by $\mathcal{W}$ and the other will be denote by μ. The measure μ will typically discrete and supported on bounded variation paths. Our aim will be to compute

$$K_{\phi}^{\mathcal{W},\mu}(s,t) = \langle \mathbb{E}_{X\sim\mathcal{W}}[S(X)_{0,1}], \mathbb{E}_{X\sim\mu}[S(X)_{0,1}] \rangle_{\phi},$$

where $S(X)$ denotes the Stratonovich signature of X. We will sometimes write $\mathbb{E}[S(\circ B)_{0,s}]$, for a Brownian motion B, in place of $\mathbb{E}_{X\sim\mathcal{W}}[S(X)_{0,1}]$ to emphasise the fact that the signature is constructed via Stratonvich calculus.

TABLE 1
Test set classification accuracy (in %) on UEA multivariate time-series data sets

Data sets	Without add-time operation			With add-time operation		
	Original	Factorial	Beta(1)	Original	Factorial	Beta(1)
ArticularyWordRecognition	81.3	80	79.7	94.3	92.3	93.3
BasicMotions	87.5	90	**100**	97.5	97.5	95
Cricket	62.5	58.3	**75**	84.7	81.9	83.3
Epilepsy	90.6	88.4	90.6	92	92	**93.5**
ERing	75.6	**78.1**	74.4	80	**87.4**	**86.3**
FingerMovements	44	**47**	**45**	49	**51**	**57**
Libras	48.9	**50**	**57.2**	66.1	65	**68.3**
NATOPS	73.9	73.3	**78.9**	90.6	88.3	**91.7**
RacketSports	69.1	68.4	67.1	78.9	78.3	**79.6**
SelfRegulationSCP1	50.5	50.5	**51.5**	70.3	**71**	68.6
UWaveGestureLibrary	74.1	73.4	**76.9**	71.9	70.3	70.6

As an initial step, we assume that γ is a fixed (deterministic) continuous path of bounded variation. We look to obtain formula for the ϕ-signature kernel of the expected Stratonovich signature of Brownian motion and γ, that is,

$$K_\phi^{\mathcal{W},\gamma}(s,t) := \langle \mathbb{E}[S(\circ B)_{0,s}], S(\gamma)_{0,t}\rangle_\phi .$$

A key idea to doing this will be to use notion of the hyperbolic development of γ, which has been used in earlier study of the signature and, in this context, was initiated by [14]. We summarise the essential background in the section below.

6.1. *Hyperbolic development.* We gather the basic notation and results. Readers seeking further details can consult [2, 14, 22]. We let $\mathbb{H}^d$ denote d-dimensional hyperbolic space realised as the hyperboloid $\{x \in \mathbb{R}^{d+1} : x * x = -1, x_{d+1} > 0\}$ endowed with the Minkowski product

$$x * y = \sum_{i=1}^d x_i y_i - x_{d+1} y_{d+1} \quad \text{for } x = (x_1, \dots, x_d, x_{d+1}) \in \mathbb{R}^{d+1}.$$

It is well known that this defines a Riemannian metric when restricted to the tangent bundle of $\mathbb{H}^d$. We let $\mathrm{d}_{\mathbb{H}^\mathrm{d}}$ denote the associated Riemannian distance function and recall that

$$\cosh \mathrm{d}_{\mathbb{H}^\mathrm{d}}(x, y) = -x * y; \tag{6.1}$$

see, for example, [3]. Define the linear map $F : \mathbb{C}^d \to \mathcal{M}_{d+1}(\mathbb{C})$ into the space of $d+1$ by $d+1$ matrices over $\mathbb{C}$ by

$$F : x \to \begin{pmatrix} 0 & x \\ x^T & 0 \end{pmatrix}. \tag{6.2}$$

Then if V is a real inner product space of dimension d and $\gamma : [a,b] \to V$ is continuous path of bounded variation then, by fixing an orthonormal basis of V, and writing γ in this basis as $\gamma = (\gamma^1, \dots, \gamma^d)$ we can uniquely solve the differential equation with linear vector fields given by

$$d\Gamma_{s,t}(u) = F(d\gamma_u)\Gamma_{s,t}(u), \quad u \in [s,t] \subset [a,b], \quad \text{with} \quad \Gamma_{s,t}(s) = I = I_{d+1}. \tag{6.3}$$

In this case, the map $\gamma|_{[s,t]} \mapsto \Gamma_{s,t}(\cdot)$ takes a path segment in V into one in the isometry group of $\mathbb{H}^d$. The resulting $\Gamma(s,t)(\cdot)$ is called the *Cartan Development of the path segment* $\gamma|_{[s,t]}$. It satisfies the multiplicative property

$$\Gamma(u,t)(t)\Gamma(s,u)(u) = \Gamma(s,t)(t), \quad s \le u \le t. \tag{6.4}$$

To simplify things, we suppress the dependence on the interval and write $\Gamma(t) := \Gamma^{\gamma}(t) := \Gamma(a,b)(t)$ for $t \in [a,b]$. It is elementary to represent Γ as the convergent series

$$\Gamma(t) = I + \sum_{n=1}^{\infty} \int_{a<t_1<\cdots<t_n<t} F(d\gamma_{t_1}) \cdots F(d\gamma_{t_n}). \tag{6.5}$$

Then letting $o = (0,\dots,0,1)^T \in \mathbb{H}^d$, we define $\sigma(t) := \Gamma(t)o$ to be the *hyperbolic development of the path* γ onto $\mathbb{H}^d$, and we write σ_γ to emphasise the dependence on γ.

A global coordinate chart for $\mathbb{H}^d$ is determined by $\mathbb{H}^d \ni m \mapsto (\eta,\rho) \in \mathbb{S}^{d-1} \times \mathbb{R}_+$ where $(\eta \sinh\rho, \cosh\rho) = m$. Using these coordinates, we define

$$\eta(t) = \eta_\gamma(t) = \eta(\sigma_\gamma(t)) \in \mathbb{S}^{d-1} \quad \text{and} \quad \rho(t) = \rho_\gamma(t) = \rho(\sigma_\gamma(t)) \in \mathbb{R}_+.$$

The following identity follows from (6.5) and (6.1):

$$\cosh\rho_\gamma(t) = \Gamma_{d+1,d+1}(t) = 1 + \sum_{n=1}^{\infty} \int_{a<t_1<\cdots<t_{2n}<t} \langle d\gamma_{t_1}, d\gamma_{t_2}\rangle \cdots \langle d\gamma_{t_{2n-1}}, d\gamma_{t_{2n}}\rangle, \tag{6.6}$$

where $\Gamma(t) = (\Gamma_{ij}(t))_{i,j=1,\dots,d+1}$. We will need to broaden this discussion to consider the development of paths after complex rescaling. To this end, if γ is as above and $z \in \mathbb{C}$ then we let $z\gamma$ denote the path in $V^{\mathbb{C}}$, the complexification of V. We will be interested in the relationship between the solution to (6.3), when γ is replaced by $z\gamma$, and the series (6.6). The following lemma identifies the structure we need.

LEMMA 6.1. *Let $\gamma : [a,b] \to V$ be a continuous path of bounded variation. For $z \in \mathbb{C}$, let $z\gamma : [a,b] \to V^{\mathbb{C}}$ be the rescaling of γ by $z \in \mathbb{C}$. Given an orthonormal basis of V, write $\gamma_t = (\gamma_t^1,\dots,\gamma_t^d) \in \mathbb{R}^d$ and $z\gamma(t) := (z\gamma_t^1,\dots,z\gamma_t^d) \in \mathbb{C}^d$ in terms of this basis. Then*

$$d\Gamma^{z\gamma}(u) = F(d(z\gamma)(u))\Gamma^{z\gamma}(u), \quad u \in [a,b], \quad \textit{with} \quad \Gamma^{z\gamma}(s) = I_{d+1} \tag{6.7}$$

has a unique solution in $\mathcal{M}_{d+1}(\mathbb{C})$, and furthermore, the entry

$$\Gamma^{z\gamma}_{d+1,d+1}(t) = 1 + \sum_{n=1}^{\infty} z^{2n} \int_{0<t_1<\cdots<t_{2n}<t} \langle d\gamma_{t_1}, d\gamma_{t_2}\rangle \cdots \langle d\gamma_{t_{2n-1}}, d\gamma_{t_{2n}}\rangle. \tag{6.8}$$

If γ is a piecewise linear path defined by the concatenation,

$$\gamma_{v_1} * \gamma_{v_2} \cdots * \gamma_{v_n} : [a,b] \to V,$$

that is, γ is such that $\gamma'_{v_i}(t) = v_i \in \mathbb{R}^d$ for $t \in (t_{i-1}, t_i)$. Then the solution to (6.7) is given explicitly by the matrix product

$$\Gamma^{z\gamma}(b) = A(v_n, \Delta_n, z)A(v_{n-1}, \Delta_{n-1}, z) \cdots A(v_1, \Delta_1, z), \tag{6.9}$$

where $\Delta_i = t_t - t_{i-1}$ and

$$A(v,\Delta,z) := I_{d+1} + \sinh(z|v|\Delta)M + (\cosh(z|v|\Delta) - 1)M^2 \tag{6.10}$$

in which

$$M = \begin{pmatrix} 0 & \tilde{v} \\ \tilde{v}^T & 0 \end{pmatrix} \in \mathcal{M}_{d+1}(\mathbb{R}) \quad \textit{with} \quad \tilde{v} = \frac{v}{|v|}.$$

PROOF. Since the ODE (6.7) is linear, there is a unique solution $\Gamma^{z\gamma}(t)$, which can be represented by equation (6.5) by replacing γ with $z\gamma$. Then equation (6.8) can be obtained by taking the last entry of this equation.

To obtain the explicit solution in the case where γ is piecewise linear path, we first assume $\gamma' = v$ on $[s, t]$. Then by using the observation that $M^3 = M$ together with equation (6.5), we have

$$\begin{aligned}\Gamma^{z\gamma}_{s,t}(t) &= I + \sum_{n=1}^{\infty} \frac{(z|v|)^{2n-1}(t-s)^{2n-1}}{(2n-1)!} M + \sum_{n=1}^{\infty} \frac{(z|v|)^{2n}(t-s)^{2n}}{(2n)!} M^2 \\ &= I + \sinh\bigl(z|v|(t-s)\bigr)M + \bigl(\cosh\bigl(z|v|(t-s)\bigr) - 1\bigr)M^2.\end{aligned}$$

In the general case, the multiplicative property (6.4) together with simple induction argument implies that the solution has the form (6.9). □

6.2. *Signature kernels and hyperbolic development.* We begin this subsection by giving a closed form of the ϕ-signature kernel $K_\phi^{\mathcal{W},\mu}(s,t)$ for the special case $\phi(k) = (\frac{k}{2})!$ based on the theory presented above.

THEOREM 6.2 (Formula for $\langle \mathbb{E}[S(\circ B)], S(\gamma)\rangle_\phi$). *Let $\phi : \mathbb{N} \cup \{0\} \to \mathbb{R}_+$ be defined by $\phi(k) = (\frac{k}{2})!$ for $k \in \mathbb{N} \cup \{0\}$. Suppose that B is a d-dimensional Brownian motion, then the expected Stratonovich signature, $\mathbb{E}[S(\circ B)_{0,s}]$, belongs to $T_\phi((V))$ for any $0 \le s < \infty$. Furthermore, if $\gamma : [0,1] \to V$ is any continuous path of bounded variation it holds that*

$$K_\phi^{\mathcal{W},\gamma}(s,t) := \bigl\langle \mathbb{E}\bigl[S(\circ B)_{0,s}\bigr], S(\gamma)_{0,t}\bigr\rangle_\phi = \cosh\bigl(\rho_{\sqrt{s/2}\gamma}(t)\bigr) \quad \textit{for all } t \textit{ in } [0,1]. \tag{6.11}$$

In this notation, $\rho_\lambda(t) := \mathrm{d}_{\mathbb{H}^{\mathrm{d}}}(o, \sigma_{\lambda\gamma}(t))$ is the distance between the hyperbolic development $\sigma_{\lambda\gamma}(t)$ of the path $\lambda\gamma(\cdot)$ from $T_o\mathbb{H}^d$ onto the d-dimensional hyperbolic space. $\mathbb{H}^d$ started at the base point $o = (0,0,\dots,1) \in \mathbb{H}^d$, and $\mathrm{d}_{\mathbb{H}^{\mathrm{d}}} : \mathbb{H}^d \times \mathbb{H}^d \to [0,\infty)$ is the Riemannian distance on $\mathbb{H}^d$.

PROOF. For the first assertion, recall that (see, e.g., Proposition 4.10 in [21])

$$\mathbb{E}\bigl[S(\circ B)_{0,s}\bigr] = \exp\left(\frac{s}{2}\sum_{i=1}^{d} e_i^2\right) = \sum_{k=0}^{\infty} \frac{s^k}{2^k k!} \sum_{i_1,\dots,i_k=1}^{d} e_{i_1}^2 \cdots e_{i_k}^2$$

so that

$$\bigl\|\mathbb{E}\bigl[S(\circ B)_{0,s}\bigr]\bigr\|_\phi^2 = \sum_{k=0}^{\infty} k! \frac{s^{2k} d^k}{2^{2k}(k!)^2} = e^{s^2 d/4} < \infty.$$

For the second assertion, we have that

$$\bigl\langle \mathbb{E}\bigl[S(\circ B)_{0,s}\bigr], S(\gamma)_{0,t}\bigr\rangle_\phi = \sum_{k=0}^{\infty} \frac{s^k}{2^k} \int_{0<t_1<\cdots<t_{2k}<t} \langle d\gamma_{t_1}, d\gamma_{t_2}\rangle \cdots \langle d\gamma_{t_{2k-1}}, d\gamma_{t_{2k}}\rangle.$$

The right-hand side of this expression equals that of (6.11); see formula (6.6). □

In the following, we give some remarks on the computation of this basic signature kernel based on the above theorem.

REMARK 6.3. We re-emphasise the key points. (1) In contrast to the earlier case of two paths, we need only solve an ODE to calculate $\langle \mathbb{E}[S(\circ B)], S(\gamma)\rangle_\phi$ and not a PDE. (2) For general γ, the ODE is known, and is determined by the linear vector fields in equation (6.3). Any ODE solver such as Runge–Kutta could in principle be used to obtain numerical solutions. (3) For the piecewise linear case, the exact solution is given in equation (6.9) as a product of matrices.

6.3. *The original kernel for expected signatures.* Theorem 6.2 gives a closed-form expression for the ϕ-signature kernel of Stratonovich expected signature of Brownian motion and the signature of a bounded variation continuous path where $\phi(k) = (\frac{k}{2})!$. As previously, we will be interested in related formulae for different signature kernels. We can obtain these formulae by using an extension of the ideas developed earlier in the paper. In the case of the original signature kernel (i.e., $\phi \equiv 1$), we can make use of the classical integral representation of the reciprocal gamma function which for integers has the form:

$$\frac{1}{k!} = \frac{1}{2\pi i} \oint_C z^{-(k+1)} e^z \, dz = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ik\theta} e^{e^{i\theta}} \, d\theta, \tag{6.12}$$

where $\oint_C$ denotes the contour integral around the unit circle traversed once anticlockwise. This is an instance of the more general formula

$$\frac{1}{\Gamma(p)} = \frac{1}{2\pi i} \oint_H z^{-p} e^z \, dz, \tag{6.13}$$

where H is a Hankel contour, which winds from $-\infty - 0i$ in the lower half-plane, anticlockwise around 0 and then back to $-\infty + 0i$ in the upper half-plane, while respecting the branch cut of the integrand along the negative real axis. The advantage of using these integral representation is twofold. First, the integrand has exponential dependence on k making it suitable to employ the techniques developed earlier in the paper. Second, the underlying numerical integration theory is well developed and the convergence rates for optimised quadrature formulae are exceedingly fast. We give some examples below but refer the reader to [31] for further details. We have the following theorem.

THEOREM 6.4. *Let $\phi \equiv 1$. Suppose B is a d-dimensional Brownian motion, then the expected Stratonovich signature, $\mathbb{E}[S(\circ B)_{0,s}]$, belongs to $T_\phi((V))$ for any $0 \le s < \infty$ and*

$$\|\mathbb{E}[S(\circ B)_{0,s}]\|_\phi^2 = \frac{1}{2\pi i} \oint_C z^{-1} e^{z + s^2 d/(4z)} \, dz, \tag{6.14}$$

where the contour C is the unit circle in $\mathbb{C}$ traversed anticlockwise. Furthermore, if γ is any continuous path of bounded variation it holds that

$$K_\phi^{\mathcal{W},\gamma}(s,t) := \langle \mathbb{E}[S(\circ B)_{0,s}], S(\gamma)_{0,t} \rangle_\phi = \frac{1}{2\pi i} \oint_C z^{-1} e^z \Gamma_{d+1,d+1}^{c_s(z)\gamma}(t) \, dz, \tag{6.15}$$

where $c_s(z) = \sqrt{s/2z} \in \mathbb{C}$ and $\Gamma_{d+1,d+1}^{c_s(z)\gamma}(t)$ is defined by the series (6.8), that is, the last diagonal entry of the matrix-valued solution to differential equation (6.7).

PROOF. Using the definition of the original signature kernel and the dominated convergence theorem to interchange the order of $\sum$ and $\oint_C$, we have

$$\begin{aligned}
\langle \mathbb{E}[S(\circ B)_{0,s}], S(\gamma)_{0,t} \rangle_\phi &= \sum_{k=0}^{\infty} \frac{1}{k!} \frac{s^k}{2^k} \int_{0<t_1<\cdots<t_{2k}<t} \langle d\gamma_{t_1}, d\gamma_{t_2} \rangle \cdots \langle d\gamma_{t_{2k-1}}, d\gamma_{t_{2k}} \rangle \\
&= \frac{1}{2\pi i} \oint_C z^{-1} e^z \\
&\quad \times \left(\sum_{k=0}^{\infty} z^{-k} \frac{s^k}{2^k} \int_{0<t_1<\cdots<t_{2k}<t} \langle d\gamma_{t_1}, d\gamma_{t_2} \rangle \cdots \langle d\gamma_{t_{2k-1}}, d\gamma_{t_{2k}} \rangle \right) dz.
\end{aligned}$$

If $c_s(z) = \sqrt{s/2z}$, then by equation (6.8), we know that

$$\sum_{k=0}^{\infty} z^{-k} \frac{s^k}{2^k} \int_{0<t_1<\cdots<t_{2k}<t} \langle d\gamma_{t_1}, d\gamma_{t_2} \rangle \cdots \langle d\gamma_{t_{2k-1}}, d\gamma_{t_{2k}} \rangle = \Gamma_{d+1,d+1}^{c_s(z)\gamma}(t),$$

which is the last entry of the solution $\Gamma^{c_{s,z}\gamma}(t)$ to ODE (6.7). The argument for the squared norm of Brownian motion, follows a similar pattern and yields

$$\|\mathbb{E}[S(\circ B)_{0,s}]\|_{\phi}^{2} = \frac{1}{2\pi i}\oint_{C} z^{-1}e^{z}\left(\sum_{k=0}^{\infty} z^{-k}\frac{s^{2k}d^{k}}{2^{2k}k!}\right)dz = \frac{1}{2\pi i}\oint_{C} z^{-1}e^{z}e^{s^{2}d/(4z)}\,dz. \qquad \square$$

Computation of the contour integrals. The implementation of the formula above demands an efficient way to approximate contour integrals of the form

$$I = \frac{1}{2\pi i}\oint_{C} e^{z}f(z)\,dz = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{e^{i\theta}} f(e^{i\theta})e^{i\theta}\,d\theta. \tag{6.16}$$

A natural approach is to apply a trapezoidal rule based on N equally spaced points on the unit circle, that is, to approximate I using

$$I_N = \frac{1}{N}\sum_{k=1}^{N} e^{z_k} f(z_k)z_k, \tag{6.17}$$

where $z_k = e^{2k\pi i/N}$. Several other methods have been proposed in Trefethen, Weideman and Schmelzer [31] for the efficient approximation of the Hankel-type contour integrals of the form

$$I = \frac{1}{2\pi i}\oint_{H} e^{z}f(z)\,dz.$$

The idea is to seek an optimal selection of contour according to the number of points in the quadrature formula. Letting $\varphi(\theta)$ be an analytic function that maps the real line $\mathbb{R}$ onto the contour H. Then the approach is to approximate

$$I = \frac{1}{2\pi i}\int_{-\infty}^{+\infty} e^{\varphi(\theta)} f\big(\varphi(\theta)\big)\varphi'(\theta)\,d\theta$$

by

$$I_N = -iN^{-1}\sum_{k=1}^{N} e^{z_k} f(z_k)w_k = -\sum_{k=1}^{N} c_k f(z_k) \tag{6.18}$$

on the finite interval $[-\pi,\pi]$ with N points, which are regularly spaced on the interval and $z_k = \varphi(\theta_k)$, $w_k = \varphi'(\theta_k)$ and $c_k = iN^{-1}e^{z_k}w_k$. The convergence rates for these optimised quadrature formulae are very fast, of order $O(3^{-N})$. Three classes of contours have been investigated in [31]:

- Parabolic contours

$$\varphi(\theta) = N\big(0.1309 - 0.1194\theta^{2} + 0.2500i\theta\big).$$

- Hyperbolic contours

$$\varphi(\theta) = 2.246N\big(1 - \sin(1.1721 - 0.3443i\theta)\big).$$

- Cotangent contours

$$\varphi(\theta) = N\big(0.5017\theta\cot(0.6407\theta) - 0.6122 + 0.2645i\theta\big).$$

Note in each case the dependence of the family on N.

The procedure for computing the kernel in equation (6.15) is first compute the function $\Gamma_{d+1,d+1}^{c_s(z)\gamma}(t)$ by utilising the explicit formula (6.9) for piecewise linear paths. By taking

$$f(z) = z^{-1}\Gamma_{d+1,d+1}^{c_s(z)\gamma}(t),$$

we can approximate the contour integral by one of the approaches described above.

6.4. *Expected signatures for general kernels.* The representation of the previous subsection can be combined with the ideas of Section 4 to obtain similar representations for $\langle \mathbb{E}[S(\circ B)], S(\gamma)\rangle_\phi$ for general ϕ satisfying the conditions of Theorem 4.11. The expression is as follows.

THEOREM 6.5. *Let μ be a finite signed Borel measure μ on $\mathbb{R}$. Suppose that $\phi : \mathbb{N} \cup \{0\} \to \mathbb{C}$ is such that*

$$\phi(k) = \int_G g(\widetilde{z})^{\alpha k} \mu(d\widetilde{z}) \in \mathbb{C} \quad \textit{for all } k \in \mathbb{N} \cup \{0\}$$

for some function $g : \mathbb{C} \to \mathbb{C}$. We assume that ϕ satisfies the conditions in Theorem 4.11, and that B is a d-dimensional Brownian motion. Then the expected Stratonovich signature, $\mathbb{E}[S(\circ B)_{0,s}]$, belongs to $T_{|\phi|}((V))$ for any $0 \le s < \infty$ and

$$\|\mathbb{E}[S(\circ B)_{0,s}]\|_\phi^2 = \frac{1}{2\pi i} \oint_C \int_G \left[z^{-1} e^z \exp\left(\frac{g(\widetilde{z})^{2\alpha} s^2 d}{4z}\right)\right] \mu(d\widetilde{z})\, dz, \tag{6.19}$$

where C is unit circle in $\mathbb{C}$ traversed anticlockwise. Furthermore, if γ is any continuous path of bounded variation it holds that

$$K_\phi^{\mathcal{W},\gamma}(s,t) := \langle \mathbb{E}[S(\circ B)_{0,s}], S(\gamma)_{0,t}\rangle_\phi = \frac{1}{2\pi i} \oint_C \int_G [z^{-1} e^z \Gamma_{d+1,d+1}^{c_{g,\alpha,s}(\widetilde{z},z)\gamma}(t)] \mu(d\widetilde{z})\, dz, \tag{6.20}$$

where $c_{g,\alpha,s}(\widetilde{z}, z) := g(\widetilde{z})^\alpha \sqrt{s/(2z)} \in \mathbb{C}$ and $\Gamma_{d+1,d+1}^{c_{g,\alpha,s}(\widetilde{z},z)\gamma}(t)$ is the series (6.8), that is, the last diagonal entry of the solution to differential equation (6.7).

PROOF. The conditions for ϕ in Theorem 4.11 and by now standards estimates allow for the steps of the proof of Theorem 6.4 to be repeated making the obvious modifications. □

As a special case, if ϕ is the moments of a random variable π, that is,

$$\phi(k) = \mathbb{E}[\pi^k] \quad \forall k \ge 0, \tag{6.21}$$

the representations are as follows.

COROLLARY 6.6. *Let the function $\phi : \mathbb{N} \cup \{0\} \to \mathbb{R}$ as defined in (6.21) and $\psi(k) = \mathbb{E}[|\pi|^k]$ such that ψ satisfies Condition 1. Suppose B is a d-dimensional Brownian motion, then the expected Stratonovich signature, $\mathbb{E}[S(\circ B)_{0,s}]$, belongs to $T_{|\phi|}((V))$ for any $0 \le s < \infty$ and*

$$\|\mathbb{E}[S(\circ B)_{0,s}]\|_\phi^2 = \frac{1}{2\pi i} \oint_C z^{-1} e^z \mathbb{E}_\pi [e^{(\pi s)^2 d/(4z)}]\, dz. \tag{6.22}$$

If γ is any continuous path of bounded variation, it holds that

$$K_\phi^{\mathcal{W},\gamma}(s,t) := \langle \mathbb{E}[S(\circ B)_{0,s}], S(\gamma)_{0,t}\rangle_\phi = \frac{1}{2\pi i} \oint_C z^{-1} e^z \mathbb{E}_\pi [\Gamma_{d+1,d+1}^{c_s(\pi,z)\gamma}(t)]\, dz, \tag{6.23}$$

where $c_s(x, z) := x\sqrt{s/(2z)} \in \mathbb{C}$ and $\Gamma_{d+1,d+1}^{c_s(x,z)\gamma}(t)$ is the series (6.8).

As an example, we recall the case $\phi(k) = \frac{\Gamma(m+1)\Gamma(k+1)}{\Gamma(k+m+1)}$ studied already in Section 4. Suppose the random variable $\pi \sim \mathrm{Beta}(1, m)$ is Beta distributed, then the moments of π are

$$\mathbb{E}[\pi^k] = \frac{B(k+1, m)}{B(1, m)} = \phi(k).$$

We then have the following.

EXAMPLE 6.7. Let $\phi(k) = \frac{\Gamma(m+1)\Gamma(k+1)}{\Gamma(k+m+1)}$ and B a d-dimensional Brownian motion. Then ϕ satisfies Condition 1. The expected Stratonovich signature, $\mathbb{E}[S(\circ B)_{0,s}]$, is well-defined and belongs to $T_\phi((V))$ for any $0 \le s < \infty$, and the squared norm

$$\|\mathbb{E}[S(\circ B)_{0,s}]\|_\phi^2 = \frac{\Gamma(m+1)}{2\pi i} \oint_C z^{-(m+1)} e^z \frac{dz}{\sqrt{1 - s^2 d / z^2}}. \tag{6.24}$$

If γ is any continuous path of bounded variation, then

$$K_\phi^{\gamma,W}(s,t) = \frac{\Gamma(m+1)}{2\pi i} \oint_C z^{-(m+1)} e^z \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \Gamma_{d+1,d+1}^{c_s(x,z)\gamma}(t) e^{-\frac{x^2}{2}} \, dx \right] dz, \tag{6.25}$$

where $c_s(x,z) = z^{-1} x \sqrt{s} \in \mathbb{C}$ and $\Gamma_{d+1,d+1}^{c_s(x,z)\gamma}(t)$ is the series (6.8).

The representations above are slightly different from Corollary 6.6 in which π should be a Beta random variable. The expressions above are obtained by the formulas below:

$$\frac{\Gamma(2k+1)}{2^k k!} = (2k-1)!! = \mathbb{E}_X[X^{2k}] \quad \text{and} \quad \frac{1}{\Gamma(2k+m+1)} = \oint_C z^{-(2k+m+1)} e^z \, dz,$$

where $X \sim N(0,1)$ is a standard normal random variable. In the point view of computation, the Gaussian quadrature for approximating the formula (6.25) is much easier than using the formula (6.23) with $\pi \sim \text{Beta}(1,m)$.

REMARK 6.8. In terms of the computation procedure, we take the signature kernel in equation (6.25) as an example. It can be calculated in three successive steps. First, for fixed z, x and s, get the exact value of $\Gamma_{d+1,d+1}^{c_s(x,z)\gamma}(t)$ by the explicit solution (6.9) to ODE (6.7) for a piecewise linear path. Second, approximate the expectation

$$\mathbb{E}_X[\Gamma_{d+1,d+1}^{c_s(X,z)\gamma}(t)] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \Gamma_{d+1,d+1}^{c_s(x,z)\gamma}(t) e^{-\frac{x^2}{2}} \, dx$$

by classical Gaussian quadrature on the whole real line. Third, approximate the contour integral using one of the methods described above. The steps are summarised schematically as follows:

$$K_\phi^{\mathcal{W},\gamma}(s,t) = \frac{\Gamma(m+1)}{2\pi i} \underbrace{\oint_C z^{-(m+1)} e^z}_{(3)\ \textit{Contour approximation}} \underbrace{\left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \overbrace{\Gamma_{d+1,d+1}^{c_s(x,z)\gamma}(t)}^{(1)\ \textit{explicit solution}} e^{-\frac{x^2}{2}} \, dx \right]}_{(2)\ \textit{Gaussian quadrature}} dz.$$

The general form (6.20) can also be computed by these three steps successively but the quadrature formula will generally be more complicated to implement than the Beta random variable case; see Section 5 for details.

7. Optimal discrete measures on paths. In the previous sections, we have introduced the ϕ-signature kernels. We described the method for the evaluation of these kernels for a pair of continuous bounded variation paths, and derived a closed-form expression for the expected signature against Brownian motion. In particular, given a finite collection of continuous bounded variation paths $\{\gamma_1, \gamma_2, \dots, \gamma_n\}$ on V and a discrete measure $\mu = \sum_{i=1}^n \lambda_i \delta_{\gamma_i}$ supported on this set we can evaluate

$$\|\mathbb{E}_{X\sim\mu}[S(X)_{0,1}]\|_\phi^2 = \sum_{i,j=1}^n \lambda_i \lambda_j K_\phi^{\gamma_i,\gamma_j},$$

and also

$$\langle \mathbb{E}_{X\sim\mathcal{W}}[S(X)_{0,1}], \mathbb{E}_{X\sim\mu}[S(X)_{0,1}]\rangle_\phi,$$

where $\mathcal{W}$ denotes the Wiener measure. This can be used to measure the similarity of using the maximum mean discrepancy distance associated with the ϕ signature kernel:

$$d_\phi^2(\mathcal{W}, \mu) = \|\mathbb{E}_{X\sim\mathcal{W}}[S(X)_{0,1}] - \mathbb{E}_{X\sim\mu}[S(X)_{0,1}]\|_\phi^2,$$

which can be used as the basis of goodness-of-fit tests to measure the similarity of μ to Wiener measure. We refer to [13] and [9] where kernels have been proposed as a way to support similar analyses.

Changing our perspective, we can also attempt to find the optimiser over some subset of measures C, that is,

$$\mu^* = \arg\min_{\mu\in C}\|\mathbb{E}_{X\sim\mathcal{W}}[S(X)_{0,1}] - \mathbb{E}_{X\sim\mu}[S(X)_{0,1}]\|_\phi^2 \tag{7.1}$$

to give the d_ϕ-best approximation to Wiener measure on C. An example in which this is tractable is when the support of μ in C is fixed to be $\{\gamma_1, \gamma_2, \dots, \gamma_n\}$ and where the set over which the optimisation is carried out is the set of probability measures with this support. In other words, C can be identified with the simplex $C_n = \{\lambda : \sum_{i=}^n \lambda_i = 1, \lambda_i \geq 0\}$. By finding this optimum, we can then compare the value $d_\phi(\mathcal{W}, \mu)$, for a given measure μ, to the optimised value $d_\phi(\mathcal{W}, \mu^*)$ to and use as a guide to whether μ is d_ϕ-close to $\mathcal{W}$ when compared to discrete measures having the same support. A closely related, although more advanced problem, is the ϕ-cubature problem of solving

$$(\mu^*, \{\gamma_i\}^*) = \arg\min_{(\mu,\{\gamma_i\})}\left\|\mathbb{E}[S(\circ B)_{0,1}] - \sum_{i=1}^n \lambda_i S(\gamma_i)_{0,1}\right\|_\phi^2,$$

which in the case where $\phi(n) = 0$ for $n \geq N$ corresponds to find a degree-N cubature formula in the sense of [21]. For N large enough, this can be minimised (not necessarily uniquely) to zero and explicit formulas for (λ_i, γ_i) are known in some case; again see [21] for more details.

7.1. *Existence and uniqueness of optimal discrete measure.* In this subsection, we consider in detail the problem described above. We give conditions on the collection $\{\gamma_1, \gamma_2, \dots, \gamma_n\}$ so that

$$L(\mu) = d_\phi^2(\mathcal{W}, \mu)$$

has a unique minimiser on the set

$$C_n = \left\{\mu = \sum_{i=1}^n \lambda_i \delta_{\gamma_i} : \lambda_i \geq 0, \lambda_1 + \dots + \lambda_n = 1\right\}.$$

In order to find the optimal discrete measure on the set of paths $\{\gamma_i\}_{i=1}^n$, we could solve the problem in equation (7.1) with constraints $\lambda_i \geq 0$ and $\sum_{i=1}^n \lambda_i = 1$. This is equivalent to solving the quadratic optimisation problem of quadratic functions with linear equality and inequality constraints given by

$$\begin{aligned} &\min_{x\in\mathbb{R}^n} \frac{1}{2}x^T K x - h^T x \\ &\text{subject to} \quad \mathbf{1}^T x = 1, \quad x \geq 0, \end{aligned} \tag{7.2}$$

where

$$K = \left(K_\phi^{\gamma_i,\gamma_j}\right)_{i,j=1,\dots,n}, \quad \text{and} \quad h = \left(K_\phi^{\gamma_1,\mathcal{W}},\dots,K_\phi^{\gamma_n,\mathcal{W}}\right)^T.$$

Existence and uniqueness of the optimal solution is guaranteed by the positive definiteness of K. Some sufficient conditions for positive definiteness can be obtained from the following lemma.

LEMMA 7.1. *The set of all signatures $\mathcal{S}$ of continuous bounded variation paths is a linearly independent subset of $T((V))$.*

PROOF. Suppose that $\{h_1,\dots,h_n\}$ is a subset of $\mathcal{S}$ and suppose that $\sum_{i=1}^n \lambda_i h_i = 0$ with not all $\lambda_i = 0$, for example, suppose that $\lambda_j \neq 0$. The vectors $h_1,\dots,h_n$ are distinct and so there exist linear functionals f_i on $T((V))$ for $i \neq j$ with $f_i(h_i) = 0$ and $f_i(h_j) = 1$. Let $p : T((V)) \to \mathbb{R}$ be the polynomial $p(x) = \prod_{i\neq j} f_i(x)$ then the linear functional L defined by the shuffle product $L = f_1 \sqcup\!\sqcup f_2 \cdots \sqcup\!\sqcup f_n$ agrees with p on $\mathcal{S}$, and hence we arrive at the contradiction

$$\lambda_j = \sum_{i=1}^n \lambda_i L(h_i) = 0.$$

$\square$

COROLLARY 7.2. *Let $\{\gamma_1,\dots,\gamma_n\}$ be a collection of continuous V-valued paths of bounded variation having distinct signatures. If $\phi : \mathbb{N}\cup\{0\} \to (0,\infty)$ satisfies Condition 1, then the matrix $K = (\langle S(\gamma_i), S(\gamma_j)\rangle_\phi)_{i,j=1,\dots,n}$ is positive definite.*

PROOF. If $0 \neq x \in \mathbb{R}^n$, then the previous proposition ensures that $\sum_{i=1}^n x_i S(\gamma_i)_{a,b} \neq 0$. Since $\|\cdot\|_\phi$ is a norm, we have

$$0 < \left\| \sum_{i=1}^n x_i S(\gamma_i)_{a,b} \right\|_\phi^2 = x^T K x$$

as required. $\square$

We now prove an existence and uniqueness theorem for the d_ϕ-closest discrete probability measure to Wiener measure, which is supported on $\{\gamma_1,\dots,\gamma_n\}$.

PROPOSITION 7.3. *Let $\{\gamma_1,\dots,\gamma_n\}$ be a collection of continuous V-valued paths of bounded variation defined over $[a,b]$ and having distinct signatures. Assume that $\phi : \mathbb{N}\cup\{0\} \to (0,\infty)$ satisfies Condition 1. Let C_n denote the n-simplex $\{\mu = (\mu_1,\dots,\mu_n) : \sum_{i=1}^n \mu_i = 1, \mu_i \geq 0\}$ so that C_n is in one-to-one correspondence with the set of probability measures supported on $\{\gamma_1,\dots,\gamma_n\}$ by the identification of μ with $\sum_{i=1}^n \mu_i \delta_{\gamma_i}$. Then there exists a unique μ^*, which minimises $d_\phi(\mu,\mathcal{W})$ over μ in C_n, that is,*

$$\mu^* = \arg\min_{\mu\in C_n} \left\| \mathbb{E}\left[S(\circ B)_{a,b}\right] - \mathbb{E}_{X\sim\mu}\left[S(X)_{a,b}\right] \right\|_\phi.$$

PROOF. It is easy to verify that the set C_n is a compact and convex set in $\mathbb{R}^n$. Since $f(x) = \frac{1}{2}x^T K x - h^T x$ is continuous on the compact set C_n, then f is bounded and attains its minimum on some points in the set C_n. That means that there exist optimal solutions $x^* \in C_n$ such that

$$f(x^*) = \min_{x\in C_n} f(x).$$

Let $m = \min_{x \in C_n} f(x)$ and $x_1^*, x_2^* \in C_n$ be two optimal solutions. Then, for any $\alpha \in [0, 1]$, we have

$$\alpha x_1^* + (1-\alpha) x_2^* \in C_n$$

and

$$m \leq f(\alpha x_1^* + (1-\alpha) x_2^*) \leq \alpha f(x_1^*) + (1-\alpha) f(x_2^*) = m.$$

Thus,

$$\frac{1}{2}(x_1^*)^T K x_2^* - \frac{1}{2} h^T (x_1^* + x_2^*) = m.$$

Since

$$f(x_1^*) = \frac{1}{2}(x_1^*)^T K x_1^* - h^T x_1^* = m \quad \text{and} \quad f(x_2^*) = \frac{1}{2}(x_2^*)^T K x_2^* - h^T x_2^* = m,$$

combining above three equations together, we have

$$(x_2^* - x_1^*)^T K (x_2^* - x_1^*) = 0.$$

Since the matrix K is positive definite on $\mathbb{R}^n$, we must have that $x_1^* = x_2^*$. So, we have concluded our proof. $\square$

REMARK 7.4. The next aim is to find the optimal measure in Theorem 7.3 and the minimised value of the objective. In some cases, this can be done explicitly. Letting f be the function in the proof, we have the following cases.

Case 1 There exists $x^* \in C_n$ such that $\nabla f(x^*) = 0$. Then the optimal solution and the value are

$$x^* = K^{-1} h \in C_n, \quad f(x^*) = -\frac{1}{2} h^T K^{-1} h.$$

Case 2 Assume that ∇f is nonvanishing on C_n. If there exists a vertex e_m of C_n such that $f(e_m) < f(e_j)$ for all $j \neq m$ and if it satisfies that

$$(K e_m - h)^T (e_i - e_m) \geq 0 \quad \forall i \in [n] := \{1, 2, \dots, n\}, \tag{7.3}$$

then the optimal solution is e_m and $f(e_m) = \frac{1}{2} e_m^T K e_m - h^T e_m$. Actually, we have

$$\begin{aligned} f(x) - f(e_m) &= \nabla f(e_m)^T (x - e_m) + \frac{1}{2}(x - e_m)^T \nabla^2 f(e_m)(x - e_m) \\ &= (K e_m - h)^T (x - e_m) + \frac{1}{2}(x - e_m)^T K (x - e_m) \\ &= (K e_m - h)^T \left(\sum_{i=1}^n \alpha_i e_i - e_m \right) + \frac{1}{2}(x - e_m)^T K (x - e_m) \\ &= \sum_{i=1}^n \alpha_i (K e_m - h)^T (e_i - e_m) + \frac{1}{2}(x - e_m)^T K (x - e_m) \\ &\geq 0, \end{aligned}$$

where $x = \sum_{i=1}^n \alpha_i e_i$ is a convex combination of vertexes of C_n. The condition (7.3) means that

$$\begin{aligned} \tilde{f}(t) &= f((1-t) e_m + t e_i) \\ &= \frac{1}{2}(e_i - e_m)^T K (e_i - e_m) t^2 + (K e_m - h)^T (e_i - e_m) t + f(e_m) \end{aligned}$$

is increasing on the interval $[0, 1]$.

If ∇f does not vanish in C_n and the conditions in case 2 of the above do not hold, then there is no explicit expression for the optimal solution and alternative numerical methods are needed to determine the minimiser. Common tools are active-set methods and interior point methods; see [35, 36] and the references therein.

8. Examples and numerical results. In this section, we give some numerical results to illustrate the usefulness of general signature kernels in measuring the similarity or alignment between a given discrete measures on paths and Wiener measure. We illustrate the use of these measures in a number of examples. As in the previous section, let $\mu = \sum_{i=1}^{n} \lambda_i \delta_{\gamma_i}$ be a discrete probability measure supported on a finite collection of continuous bounded variation paths $\gamma : [0, 1] \to V$ and denote the Wiener measure on $\mathcal{W}$. A plausible measure of the alignment between these two expected signatures is

$$\cos \angle_{\phi}(\mu, \mathcal{W}) := \frac{\langle \mathbb{E}_{X\sim\mu}[S(\circ X)_{0,1}], \mathbb{E}_{X\sim\mathcal{W}}[S(X)_{0,1}]\rangle_{\phi}}{\|\mathbb{E}_{X\sim\mu}[S(\circ X)_{0,1}]\|_{\phi}\|\mathbb{E}_{X\sim\mathcal{W}}[S(X)_{0,1}]\|_{\phi}}. \tag{8.1}$$

It follows from our earlier discussion that $\cos \angle_{\phi}(\mu, \mathcal{W}) \in [0, 1]$. A justification for this quantity measuring the alignment of the measures μ and $\mathcal{W}$, rather than just their expected signatures is that for any given pair of measures ν_1 and ν_2 on a space of (rough) paths it holds that $\cos \angle_{\phi}(\nu_1, \nu_2) = 1$ if and only if there exists $\lambda \in \mathbb{R}$ with

$$\mathbb{E}_{X\sim\nu_1}\big[S(\circ X)_{0,1}\big] = \lambda \mathbb{E}_{X\sim\nu_2}\big[S(\circ X)_{0,1}\big].$$

The fact that $\lambda = 1$, and hence that the expected signatures coincide, follows by interpreting this equality under the projection $\pi_0 : T((V)) \to \mathbb{R}$. Another quantity we use is the MMD distance

$$d_{\phi}(\mu, \mathcal{W}) = \big\|\mathbb{E}_{X\sim\mu}\big[S(\circ X)_{0,1}\big] - \mathbb{E}_{X\sim\mathcal{W}}\big[S(X)_{0,1}\big]\big\|_{\phi}, \tag{8.2}$$

which we have already discussed extensively.

8.1. *Discrete measures on Brownian paths.* In Section 7, we proved the existence of a unique optimal probability measure μ^* supported on $\{\gamma_1, \dots, \gamma_n\}$ such that

$$\mu^* = \arg\min_{\mu\in C_n} \big\|\mathbb{E}\big[S(\circ B)_{0,1}\big] - \mathbb{E}^{\mu}\big[S(\gamma)_{0,1}\big]\big\|_{\phi}^2.$$

We now present an example in which $\{\gamma_1, \dots, \gamma_n\}$ is obtained as the piecewise linear interpolation of n i.i.d. discretely-sampled Brownian paths. We consider two cases for ϕ:

(1) $\phi(k) = (\frac{k}{2})!$ for $n \in \mathbb{N} \cup \{0\}$. We refer to the resulting ϕ-signature kernel, somewhat inexactly, as the *the factorially-weighted signature kernel*.
(2) The original signature kernel $\phi(k) \equiv 1$.

EXAMPLE 8.1. We randomly sample n i.i.d. Brownian motion paths in $\mathbb{R}^d$. Each path is sampled over the time interval $[0, 1]$ on an equally-spaced partition $0 = t_1 < t_2 < \cdots < t_m = 1$ with $t_{j+1} - t_j = \frac{1}{m-1}$. We denote the resulting finite set piecewise linearly interpolated Brownian sample paths as

$$\mathcal{S}(n, m, d) = \{B_i\}_{i=1}^{n} \quad \text{with} \quad B_i = \big\{B_i(t_j) \in \mathbb{R}^d\big\}_{j=1}^{m}.$$

Figures 2 and 3 display the alignment $\cos \angle_{\phi}(\mu^*, \mathcal{W})$ and the similarity $d_{\phi}(\mu^*, \mathcal{W})$ for the optimal discrete probability measure supported on $\mathcal{S}(n, m, d)$, in which the number of sample paths $n = 10$ and the observation points $m = 10$ are fixed and the dimension d is varied over the range 2 to 6. We run 400 independent experiments for each d, that is, we generate 400 independent samples of the sets $\mathcal{S}(n, m, d)$ for each dimension d. Each set $\mathcal{S}(n, m, d)$ has an

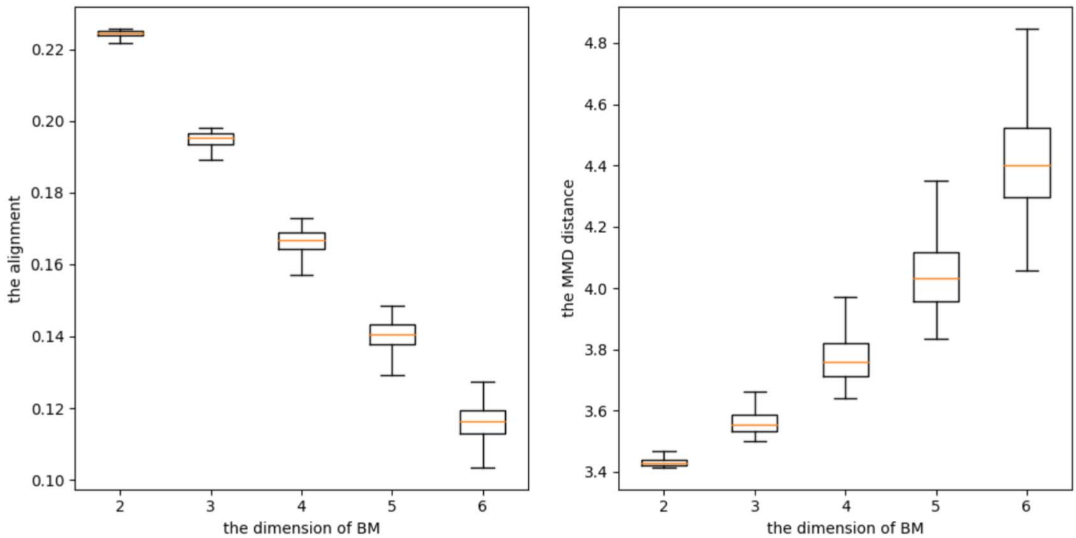

FIG. 2. *Boxplots of the factorially-weighted signature kernel*. (*a*) *The left panel shows the distribution of the values of the alignment* $\cos \angle_{\phi}(\mu^*, \mathcal{W})$ *of the optimal measure and the Wiener measure across* 400 *samples*. *The* x*-axis is the dimension of the Brownian motion*, *and the* y*-axis is the value of the alignment*. (*b*) *The right panel shows the same for the MMD distance* $d_{\phi}(\mu^*, \mathcal{W})$.

optimal measure associated with it, which we compute. The boxplots in Figures 2 and 3 show the median, range and interquartile range of the values of the alignment and the similarity of the optimal discrete measures over these 400 samples. Qualitatively, we can see from both quantities that they show dependence on the dimension of the state space, with the alignment decreasing and the dis-similarity increasing w.r.t. the dimension. We can also compare the results using the two different ϕ-signature kernels with the original signature kernel showing the same behaviour w.r.t. the dimension having a persistently higher level of alignment than under the factorially-weighted signature kernel across all of the dimensions considered.

8.2. *Examples using cubature formulae*. In the paper [21], Lyons and Victoir studied cubature on Wiener space. Let $C_{bv}([0, T], V)$ be a subset of Wiener space made of bounded

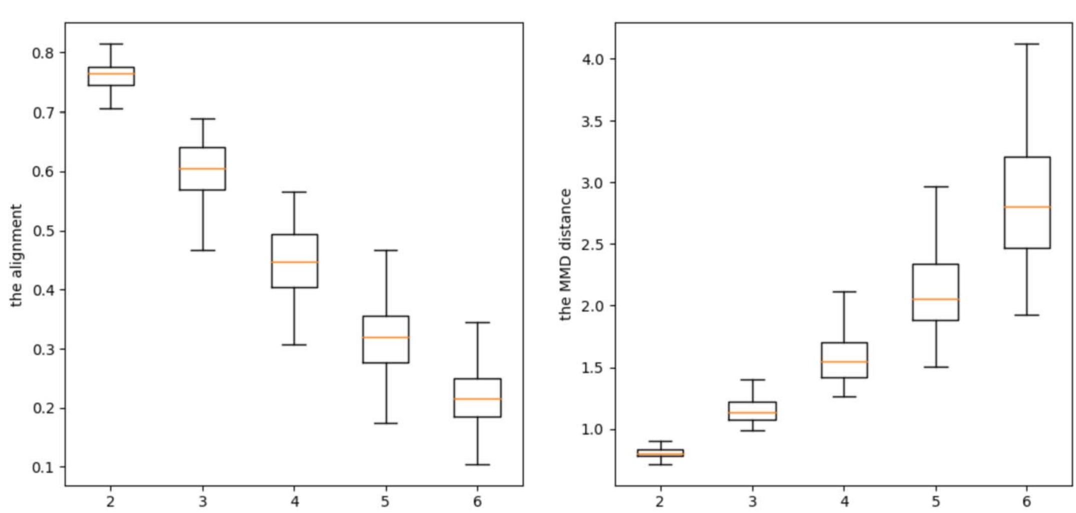

FIG. 3. *The optimal measure under the original signature kernel*.

variation paths. We say that the paths $\gamma_1, \dots, \gamma_n \in C_{bv}([0, T], V)$ and the positive weights $\lambda_1, \dots, \lambda_n$ define a cubature formula on Wiener space of degree m at time T if

$$\mathbb{E}[S^I(\circ B)_{0,T}] = \sum_{j=1}^{n} \lambda_j S^I(\gamma_j)_{0,T}$$

for all $I \in \mathcal{A}_m := \{I = (i_1, \dots, i_k) : k \leq m\}$ with $m \in \mathbb{N}$.

Cubature on Wiener space can be an effective way to develop high-order numerical schemes for high-dimensional stochastic differential equations and parabolic partial differential equations; see [21]. In Section 5 of [21], the authors also construct an explicit cubature formula of degree 5 for 2-dimensional Brownian motion. The reader can find formulas of these cubature paths and measure in Tables 2 and 3 in the same reference.

In this subsection, we analyse the results for a family of ϕ-signature kernels on three discrete probability measures supported on this collection of cubature paths. We consider the cubature weights themselves, the empirical measure of the sample (i.e., where they are equally weighted) and the optimal measure obtained from Section 7. In Figure 4, we show the similarity of these discrete measures and the Wiener measure under the family of Beta-weighted signature kernels given by

$$\phi(k) = \frac{\Gamma(m+1)\Gamma(k+1)}{\Gamma(k+m+1)} \tag{8.3}$$

for various values of m in the weight ϕ (shown along the horizontal axis).

The plot on the left panel of Figure 4 shows that as the parameter m increases these three distances first increase fast and then gradually go down. We see that the distance of the optimal measure and the Wiener measure is smallest and the distance of the empirical measure is much larger than the distance of cubature measure. The right panel shows the ratio of the distance of optimal measure and the distance of cubature measure for different choices of m.

8.3. *Applications in signal processing*. The alignment in equation (8.1) and the similarity in equation (8.2) defined by the ϕ-signature kernel give us a way of determining how large a given discrete measure is different to the Wiener measure. We can use these quantities to

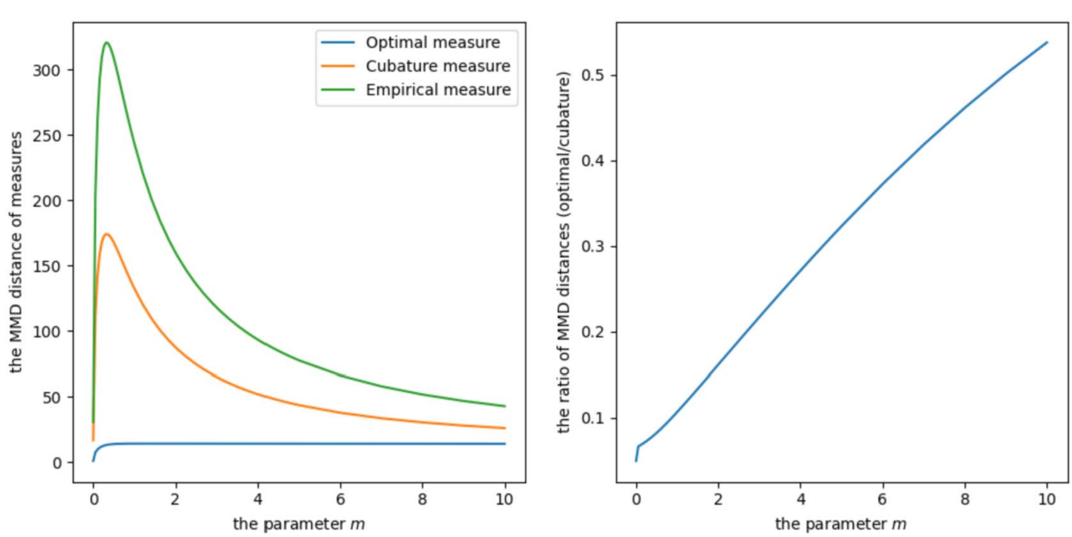

FIG. 4. *The similarity under a family of Beta-weighted signature kernels. The left panel is the plot of the distance of these discrete measures and the Wiener measure plotted against different values of m on the horizontal axis. The right panel plots the ratio of the optimal distance and the cubature distance.*

measure deviation of a discrete measure from a reference measure (i.e., the Wiener measure here). A natural application of these methods in signal processing is to mitigate/detect the (additive) contamination of white noise under different types of perturbation.

The examples studied here are motivated by an attempt to study radio frequency interference (RFI) in the radio astronomy. In this setting, astronomers would like to obtain high-resolution sky images of an interested astrophysical object using measurements from an array of antennas (e.g., the Karl G. Jansky Very Large Array (VLA), etc.). The observation is called visibility $V_{ij}(t, v, p)$, where ij is an antenna pair, t is the time integration, v is the frequency and p is the polarization. Usually, the visibility would be contaminated by thermal noise and radio frequency interference (RFI). So, the observation data from an interferometer can be broken down into three components: the astrophysical sky signals, thermal noise and RFI. The first component is slowly varying, which can be removed in the observation data by the sky-subtraction method (see, e.g., [34]). The RFI signal is usually much stronger than thermal noise but is also sometimes ultra-faint. For different antennas, the RFI contamination is systematic and thermal noise can be assumed to be independent. In order to obtain a high-resolution image, the first step is to design some methods to identify and then, if possible, to remove the RFI component of the observation.

We consider two idealised types of RFI contamination. The first is by simple superposition with a sine wave of a fixed single frequency and a given amplitude and phase, so that the interference is narrow-band but persistent over time. The second will be to consider a short duration spike, as modelled in the paper by Davis and Monroe [11] in the univariate setting, in which the Brownian signal undergoes a perturbation at a uniformly distributed random time to give

$$B(t) + \epsilon\sqrt{(t - U)^+}. \tag{8.4}$$

We again compare the use of two ϕ-signature kernels. The factorially-weighted signature kernel and the original signature kernel.

EXAMPLE 8.2. Working in d-dimensions, we take a path of the form

$$X_i^{(j)}(t) = B_i^{(j)}(t) + \epsilon \sin(2\pi\nu t - \phi_i^{(j)}), \quad j = 1, 2, \ldots, d,$$

where the frequency ν is fixed, the phase shifts are $\phi_i^{(j)}$ and ϵ denotes a (small) fixed amplitude. Let a finite collection of sample paths on time interval [0, 1] as

$$\mathcal{S}(n, m, d) = \{X_i\}_{i=1}^n \quad \text{where } X_i = \{(X_i^{(1)}(t_j), \ldots, X_i^{(d)}(t_j)) \in \mathbb{R}^d\}_{j=1}^m.$$

In Figures 5 and 6, we fixed $(n, m, d) = (10, 10, 2)$, ϵ from [0, 1] and the frequency $\nu \in \{2, 3\}$. We run 100 collections of paths $\mathcal{S}(n, m, d)$ for each ϵ and frequency ν. The figures show the deviation of the alignment and the similarity of the optimal measure (the empirical measure, resp.) and the Wiener measure, in which the middle line is the median of the alignment or the similarity, respectively, and the shadow represents the range from the lower quartile to the upper quartile. We generate 100 experiments for each ϵ. The figures show that the alignment decreases very fast to a low level and the dis-similarity increases very quickly as ϵ becomes large for both the optimal measure and the empirical measure. At larger frequencies ν, the alignment (dis-similarity) decays (grows) more rapidly.

Finally, we present an example based on the construction in the paper of Davis and Monroe [11] mentioned earlier. Here, the interference is characterised by a sudden high energy spike at a uniform random time.

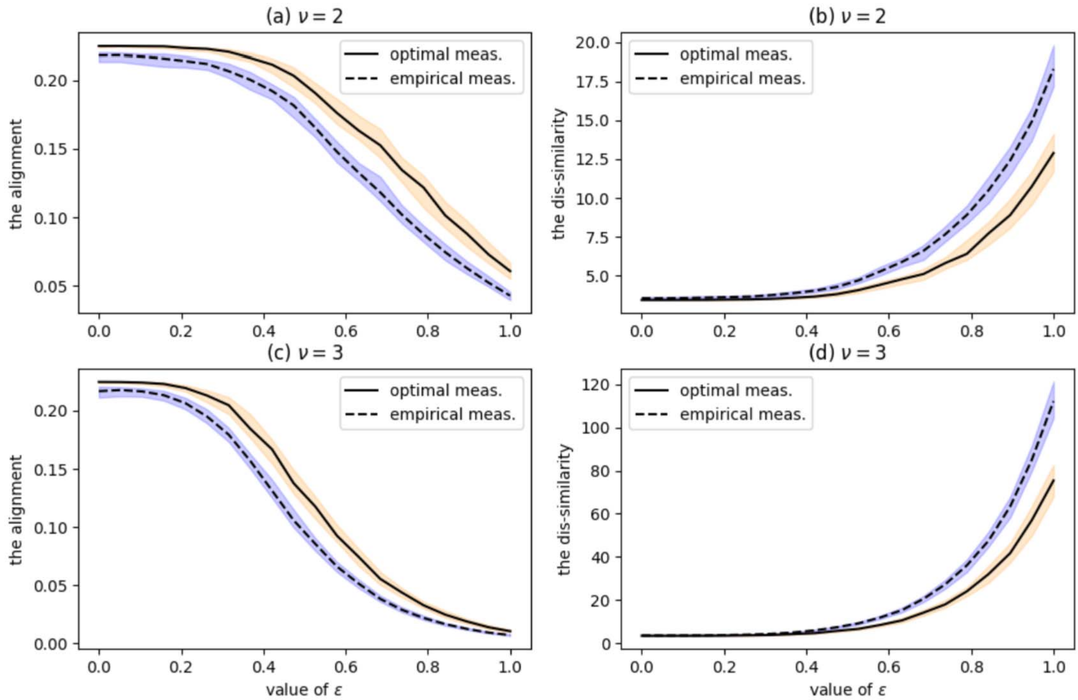

FIG. 5. *The case for the factorially-weighted signature kernel. (a) and (c) show similarities of discrete measures and the Wiener measure where the horizontal is the value of ϵ and vertical axis is the value of alignment. (b) and (d) show similarities of discrete measures and the Wiener measure. The solid line is for the optimal measure while the dashed line is for the empirical measure. The upper panel is for the frequency $\nu = 2$ and the lower is for $\nu = 3$.*

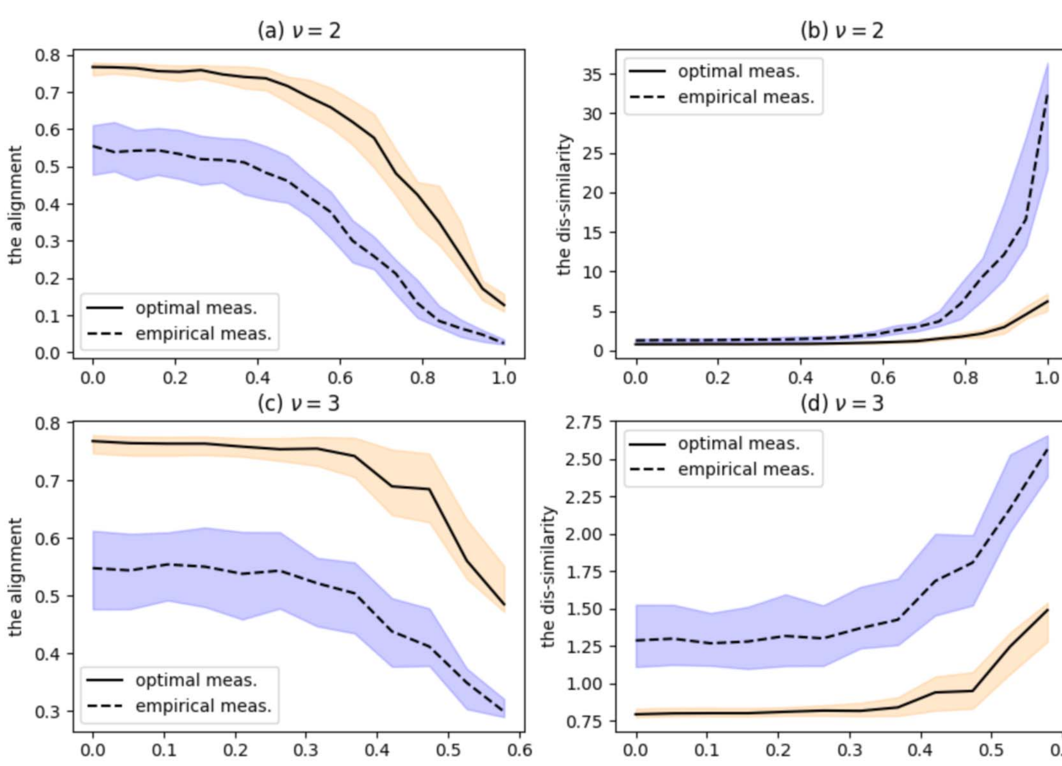

FIG. 6. *The same example under the original signature kernel.*

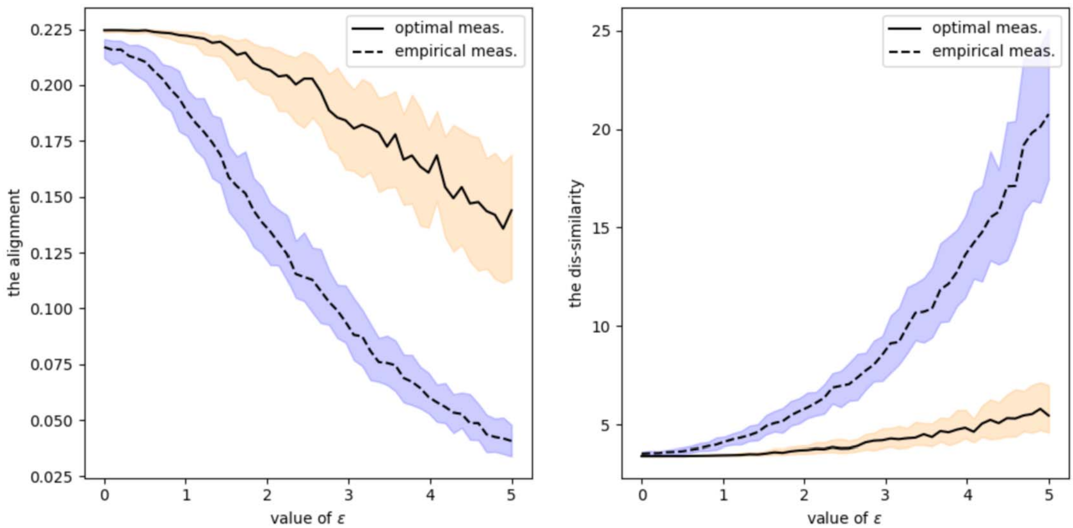

FIG. 7. *The case for the factorially-weighted signature kernel*. (*a*) *The left panel shows the alignment of discrete measures and the Wiener measure for each* ϵ *taken from* [0, 1] *where the* x*-axis is the value of* ϵ *and the* y*-axis is the value of alignment*. (*b*) *The right panel shows the similarity of discrete measures and the Wiener measure as in* (*a*). *The solid line is for the optimal measure while the dash line is for the empirical measure.*

EXAMPLE 8.3. We define

$$X_i^{(j)}(t) = B_i^{(j)}(t) + \epsilon\sqrt{(t - U_i)^+}, \quad j = 1, 2, \dots, d,$$

where U_i is uniformly distributed in $[0, 1]$, the time interval $t \in [0, 1]$ and $x^+ = \max\{0, x\}$. We denote a finite collection of these paths as

$$\mathcal{S}(n, m, d) = \{X_i\}_{i=1}^n \quad \text{where } X_i = \{(X_i^{(1)}(t_j), \dots, X_i^{(d)}(t_j)) \in \mathbb{R}^d\}_{j=1}^m.$$

In Figures 7 and 8, the parameters $(n, m, d) = (10, 10, 2)$ are fixed and ϵ is taken from $[0, 5]$. We run 100 independent experiments for each ϵ. The plots are like ones in the above example. The middle line is the median of the alignment (the similarity, resp.) and the shadow is the

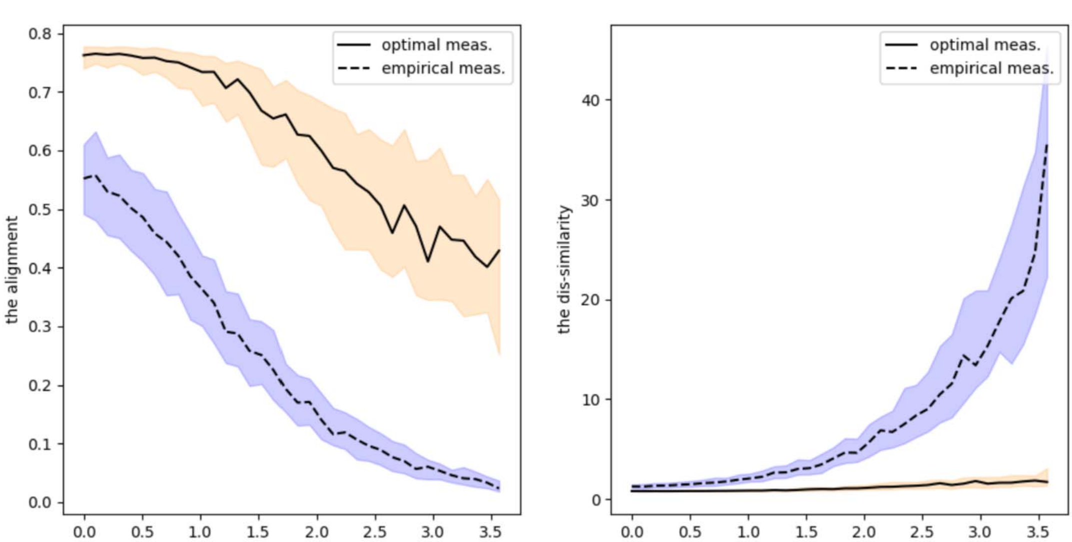

FIG. 8. *The same example under the original signature kernel.*

range from the lower quartile to the upper quartile of the alignment (the similarity, resp.) for the 100 collections of sample paths. We can see from these figures that the alignment (the dissimilarity, resp.) is decreasing (increasing, resp.) as ϵ increases, as one would expect. From the point view of RFI mitigation, the alignment of the empirical measure is more relevant than that of the optimal measure; the the values for the alignment with the optimal measure are included for comparison. It is reasonable that as the strength ϵ is large the empirical measure is less similar w.r.t. the Wiener measure than the optimal measure. The alignment of the empirical measure decays faster than that of optimal measure in our experiments. This suggests potential uses for building a method for the identification of RFI based on a threshold for the alignment of the empirical measure. The preliminary results here for instance suggest that a threshold of alignment of 0.2 under the factorially-weighted signature kernel could be used in this example.

Acknowledgements. The authors would like to thank Cris Salvi and Bojan Nikolic, respectively, for discussions related to the signature kernel and the problem of RFI mitigation in radio astronomy. The authors are also grateful to the anonymous referees for their helpful comments on the first draft of this paper.

Funding. The work of all three authors was supported by the EPSRC Programme Grant EP/S026347/1. Terry Lyons was funded in part by the EPSRC [Grant No. EP/S026347/1], in part by The Alan Turing Institute under the EPSRC Grant EP/N510129/1, the Data Centric Engineering Programme (under the Lloyd's Register Foundation Grant G0095), the Defence and Security Programme (funded by the UK Government) and the Office for National Statistics & The Alan Turing Institute (strategic partnership) and in part by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA).

REFERENCES

[1] ABRAMOWITZ, M. and STEGUN, I. A. (1964). *Handbook of Mathematical Functions*: *With Formulas*, *Graphs*, *and Mathematical Tables*. *National Bureau of Standards*: *Applied Mathematics Series* **55**. U.S. Government Printing Office, Washington, D.C. MR0167642

[2] BOEDIHARDJO, H. and GENG, X. (2019). Tail asymptotics of the Brownian signature. *Trans*. *Amer*. *Math*. *Soc*. **372** 585–614. MR3968780 https://doi.org/10.1090/tran/7683

[3] CANNON, J. W., FLOYD, W. J., KENYON, R. and PARRY, W. R. (1997). Hyperbolic geometry. In *Flavors of Geometry*. *Math*. *Sci*. *Res*. *Inst*. *Publ*. **31** 59–115. Cambridge Univ. Press, Cambridge. MR1491098

[4] CARMELI, C., DE VITO, E., TOIGO, A. and UMANITÀ, V. (2010). Vector valued reproducing kernel Hilbert spaces and universality. *Anal*. *Appl*. (*Singap*.) **8** 19–61. MR2603770 https://doi.org/10.1142/S0219530510001503

[5] CASS, T. and TURNER, W. F. (2022). Topologies on unparameterised path space. Available at arXiv:2206.11153.

[6] CHEN, K.-T. (1954). Iterated integrals and exponential homomorphisms. *Proc*. *Lond*. *Math*. *Soc*. (3) **4** 502–512. MR0073174 https://doi.org/10.1112/plms/s3-4.1.502

[7] CHEN, K.-T. (1958). Integration of paths—a faithful representation of paths by non-commutative formal power series. *Trans*. *Amer*. *Math*. *Soc*. **89** 395–407. MR0106258 https://doi.org/10.2307/1993193

[8] CHEVYREV, I. and LYONS, T. (2016). Characteristic functions of measures on geometric rough paths. *Ann*. *Probab*. **44** 4049–4082. MR3572331 https://doi.org/10.1214/15-AOP1068

[9] CHEVYREV, I. and OBERHAUSER, H. (2022). Signature moments to characterize laws of stochastic processes. Available at arXiv:1810.10971.

[10] CHWIALKOWSKI, K., STRATHMANN, H. and GRETTON, A. (2016). A kernel test of goodness of fit. In *Proceedings of the* 33*rd International Conference on Machine Learning*, *Vol*. 48, 2606–2615. PMLR.

[11] DAVIS, B. and MONROE, I. (1984). Randomly started signals with white noise. *Ann*. *Probab*. **12** 922–925. MR0744249

[12] FAWCETT, T. (2003). Problems in stochastic analysis: Connections between rough paths and non-commutative harmonic analysis Ph.D. thesis, Univ. Oxford.

[13] GRETTON, A., BORGWARDT, K. M., RASCH, M. J., SCHÖLKOPF, B. and SMOLA, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13** 723–773. MR2913716

[14] HAMBLY, B. and LYONS, T. (2010). Uniqueness for the signature of a path of bounded variation and the reduced path group. *Ann. of Math.* (2) **171** 109–167. MR2630037 https://doi.org/10.4007/annals.2010.171.109

[15] HOFMANN, T., SCHÖLKOPF, B. and SMOLA, A. J. (2008). Kernel methods in machine learning. *Ann. Statist.* **36** 1171–1220. MR2418654 https://doi.org/10.1214/009053607000000677

[16] KASHIMA, H., TSUDA, K. and INOKUCHI, A. (2003). Marginalized kernels between labeled graphs. In *Proceedings of the* 20*th International Conference on Machine Learning* (*ICML*-03) 321–328.

[17] KIRÁLY, F. J. and OBERHAUSER, H. (2019). Kernels for sequentially ordered data. *J. Mach. Learn. Res.* **20** Paper No. 31, 45. MR3948071

[18] LEMERCIER, M., SALVI, C., DAMOULAS, T., BONILLA, E. V. and LYONS, T. Distribution regression for sequential data. Available at arXiv:2006.05805.

[19] LESLIE, C., ESKIN, E. and NOBLE, W. S. (2001). The spectrum kernel: A string kernel for SVM protein classification. *Pac. Symp. Biocomput.* **2002** 564–575.

[20] LODHI, H., SAUNDERS, C., SHAWE-TAYLOR, J., CRISTIANINI, N. and WATKINS, C. (2002). Text classification using string kernels. *J. Mach. Learn. Res.* **2** 419–444.

[21] LYONS, T. and VICTOIR, N. (2004). Cubature on Wiener space. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **460** 169–198. MR2052260 https://doi.org/10.1098/rspa.2003.1239

[22] LYONS, T. J. and XU, W. (2017). Hyperbolic development and inversion of signature. *J. Funct. Anal.* **272** 2933–2955. MR3608657 https://doi.org/10.1016/j.jfa.2016.12.024

[23] SALVI, C., CASS, T., FOSTER, J., LYONS, T. and YANG, W. (2021). The signature kernel is the solution of a Goursat PDE. *SIAM J. Math. Data Sci.* **3** 873–899. MR4309861 https://doi.org/10.1137/20M1366794

[24] SCHMELZER, T. and TREFETHEN, L. N. (2007). Computing the gamma function using contour integrals and rational approximations. *SIAM J. Numer. Anal.* **45** 558–571. MR2300287 https://doi.org/10.1137/050646342

[25] SCHÖLKOPF, B., SMOLA, A. J. and BACH, F. (2002). *Learning with Kernels*: *Support Vector Machines*, *Regularization*, *Optimization*, *and Beyond*. MIT press.

[26] SHIZGAL, B. (1981). A Gaussian quadrature procedure for use in the solution of the Boltzmann equation and related problems. *J. Comput. Phys.* **41** 309–328. MR0626614 https://doi.org/10.1016/0021-9991(81)90099-1

[27] SRIPERUMBUDUR, B. K., FUKUMIZU, K. and LANCKRIET, G. R. G. (2011). Universality, characteristic kernels and RKHS embedding of measures. *J. Mach. Learn. Res.* **12** 2389–2410. MR2825431

[28] STEEN, N. M., BYRNE, G. D. and GELBARD, E. M. (1969). Gaussian quadratures for the integrals $\int_0^\infty \exp(-x^2) f(x)\,dx$ and $\int_0^b \exp(-x^2) f(x)\,dx$. *Math. Comp.* **23** 661–671. MR0247744 https://doi.org/10.2307/2004395

[29] STEINWART, I. (2002). On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.* **2** 67–93. MR1883281 https://doi.org/10.1162/153244302760185252

[30] SÜLI, E. and MAYERS, D. F. (2003). *An Introduction to Numerical Analysis*. Cambridge Univ. Press, Cambridge. MR2006500 https://doi.org/10.1017/CBO9780511801181

[31] TREFETHEN, L. N., WEIDEMAN, J. A. C. and SCHMELZER, T. (2006). Talbot quadratures and rational approximations. *BIT* **46** 653–670. MR2265580 https://doi.org/10.1007/s10543-006-0077-9

[32] TSUDA, K., KIN, T. and ASAI, K. (2002). Marginalized kernels for biological sequences. *Bioinformatics* **18** S268–S275.

[33] VAPNIK, V. (1998). The support vector method of function estimation. In *Nonlinear Modeling* 55–85. Springer, Berlin.

[34] WILENSKY, M. J., MORALES, M. F., HAZELTON, B. J., BARRY, N. and ROY, S. (2019). Absolving the SSINS of precision interferometric radio data: A new technique for mitigating ultra-faint radio frequency interference. *Publ. Astron. Soc. Pac.* **131** 114507.

[35] WONG, E. (2011). Active-set methods for quadratic programming. Ph.D. thesis, Univ. California, San Diego. MR2890183

[36] WRIGHT, S. J. (1987). *Primal-Dual Interior-Point Methods*. SIAM, Philadelphia, PA.