Signature Moments to Characterize Laws of Stochastic Processes

Ilya Chevyrev   ICHEVYREV@GMAIL.COM
School of Mathematics
University of Edinburgh
Edinburgh, EH9 3FD, United Kingdom

Harald Oberhauser   OBERHAUSER@MATHS.OX.AC.UK
Mathematical Institute
University of Oxford
Oxford, OX2 6GG, United Kingdom

Editor: Kenji Fukumizu

Abstract

The sequence of moments of a vector-valued random variable can characterize its law. We study the analogous problem for path-valued random variables, that is stochastic processes, by using so-called robust signature moments. This allows us to derive a metric of maximum mean discrepancy type for laws of stochastic processes and study the topology it induces on the space of laws of stochastic processes. This metric can be kernelized using the signature kernel which allows to efficiently compute it. As an application, we provide a non-parametric two-sample hypothesis test for laws of stochastic processes.

Keywords: stochastic processes, maximum mean discrepancy, kernel mean embedding, signature kernel, signature features, rough paths, hypothesis testing

1. Introduction

Many inference tasks about data in a topological space $\mathcal{X}$ involves

(A) making inference about a function $f \in \mathcal{F}$ from a class $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$,

(B) making inference about the probability measure μ from which the data was sampled.

A common approach is to use a "feature map", $\Phi : \mathcal{X} \to E$, that maps $\mathcal{X}$ into a linear space E which is rich enough so that both tasks, (A) and (B), can be addressed by linear methods in the "feature space" E.

Example 1 (vector-valued data) *If $\mathcal{X} \subset \mathbb{R}^d$ is compact, the map*

$$\Phi : \mathcal{X} \to \prod_{m \geq 0} (\mathbb{R}^d)^{\otimes m}, \quad x \mapsto \Big(\frac{x^{\otimes m}}{m!} \Big)_{m \geq 0} \tag{1}$$

has the property that

(a) any $f \in \mathcal{F} \equiv C(\mathcal{X}, \mathbb{R}^d)$ can be arbitrary well-approximated by a linear functional of Φ,

(b) *the map $\mu \mapsto \mathbb{E}[\Phi(x)] \in \prod_{m \geq 0} (\mathbb{R}^d)^{\otimes m}$, which maps a probability measure μ on $\mathcal{X}$ to a sequence of tensors, is injective.*

Here, (a) follows directly from the Stone–Weierstrass theorem which guarantees that linear combinations of monomials approximate continuous functions arbitrary well; (b) is the well-known fact that the moments of a bounded random variable characterize its law.

This article studies these two questions, (A) and (B), for the case that $\mathcal{X}$ is a space of paths,

$$\mathcal{X} \subset \bigcup_t C([0, t], \mathbb{R}^d) \ .$$

This includes the case of sequence-valued data since any sequence of vectors $x(t_1), \ldots, x(t_L) \in \mathbb{R}^d$ can be identified as a path $x \in C([0, t_L], \mathbb{R}^d)$ by piecewise linear interpolation between the points $x(t_1), \ldots, x(t_L)$. However, first understanding the general continuous time case guarantees that the resulting approach is robust and well-defined for sequence-valued data in the high-frequency limit when the time-discretization gets finer; see also Remark 3. Moreover, this covers important classes of processes such as stochastic differential equations or semimartingales and allows us to use tools from stochastic analysis. By using embeddings into linear spaces, such as through reproducing kernels, we are further able to handle paths evolving in general topological spaces.

Path-valued data. A classical theme from stochastic analysis is that for a path $x = (x(t))_{t \in [0,T]}$, the "signature" map

$$x \mapsto \left(\int dx^{\otimes m} \right)_{m \geq 0} \in \prod_{m \geq 0} (\mathbb{R}^d)^{\otimes m} \tag{2}$$

is an injection up to tree-like equivalence[1] and the tensor $\int dx^{\otimes m} \in (\mathbb{R}^d)^{\otimes m}$ behaves in a precise algebraic sense analogous to a monomial $\frac{y^{\otimes m}}{m!} \in (\mathbb{R}^d)^{\otimes m}$ of degree m of a vector $y \in \mathbb{R}^d$; here we use the shorthand notation $\int dx^{\otimes m}$ for the iterated integral

$$\int dx^{\otimes m} := \int_{0 < t_1 < \cdots < t_m < T} dx(t_1) \otimes \cdots \otimes dx(t_m) \tag{3}$$

$$:= \int_{0 < t_1 < \cdots < t_m < T} \dot{x}(t_1) \otimes \cdots \otimes \dot{x}(t_m) dt_1 \cdots dt_m \in (\mathbb{R}^d)^{\otimes m}, \tag{4}$$

where $\dot{x}(t) := (\frac{dx^1(t)}{dt}, \ldots, \frac{dx^d(t)}{dt}) \in \mathbb{R}^d$. The iterated integrals are defined through either classical Riemann-Stieltjes integrals, or as stochastic or rough path integrals if x is not smooth. It is then a natural question whether the expected value of these iterated integrals characterizes the law of random paths, analogous to Example 1.

Example 2 *If $Y \sim N(0, 1)$, the standard normal on $\mathbb{R}$, then*

$$\left(\mathbb{E}\left[\frac{1}{m!} Y^{\otimes m} \right] \right)_{m \geq 0} = \exp\left(\frac{1}{2} e_1 \otimes e_1 \right) \in \prod_{m \geq 0} \mathbb{R}^{\otimes m},$$

---

1. Tree-like equivalence is a very useful equivalence relation on pathspace; e.g. it identifies paths up to a time-change. From an analytic point of view, tree-like equivalence is the analogous notion of Lebesgue almost sure equivalence of sets in $\mathbb{R}^d$ on pathspace. We give a precise definition in Section B.1.

where e_1 is the unit basis vector of $\mathbb{R}$ and $\exp$ denotes the natural generalisation of the exponential map, see Appendix A. The classical Carleman's condition implies that this moment sequence uniquely determines the law of Y. A Brownian motion $X = (X_t)_{t \in [0,T]}$ in $\mathbb{R}^d$ is arguably the natural infinite-dimensional analogue of the normal distribution; indeed one has the analogous formula (Fawcett, 2003)

$$\Big(\mathbb{E}\Big[\int dX^{\otimes m}\Big]\Big)_{m \geq 0} = \exp\Big(\frac{T}{2}\sum_{i=1}^{d} e_i \otimes e_i\Big) \in \prod_{m \geq 0}(\mathbb{R}^d)^{\otimes m}\,, \tag{5}$$

where $\{e_1, \ldots, e_d\}$ is an orthonormal basis of $\mathbb{R}^d$ and the stochastic integrals are in the Stratonovich sense. It was shown in (Chevyrev and Lyons, 2016) that (5) *uniquely determines the law of Brownian motion (up to tree-like equivalence).*

Example 2 motivates the signature of a path (resp. expected signature of a stochastic process) as a generalisation of polynomials (resp. moments) to path-valued data. This viewpoint is well-known in stochastic analysis but has become increasingly popular in machine learning; we refer to (Chevyrev and Kormilitzin, 2016) for a short overview and to (Lyons et al., 2007) for a concise mathematical introduction. We emphasize that such sequences of iterated integrals go at least back to the 20th century and certain aspects of it are studied by various communities in algebra, topology, and control theory (for example, the terms time-ordered exponential, chronological calculus, and Chen-Fliess series all refer to essentially the same object that we call signature; see (Kiraly and Oberhauser, 2019, Section 1.4(iii)) for such references).

Remark 1 *It is natural to ask that, if Example 2 generalizes the moment sequence of Example 1 from vector-valued to path-valued random variables (i.e. the moment generating function), then what is the natural generalization of the Fourier transform (i.e. the characteristic function)? Indeed, Chevyrev and Lyons (2016) introduce and study such a "Fourier transform", but the resulting objects are quite abstract involving unitary representations of the signature group; in particular it is not clear how to turn this theory into computable features or kernels. Nevertheless, we believe this to be an interesting research question since already for vector-valued data, both approaches—moment sequences and Fourier transforms—are useful, although arguably moment sequences are more popular for inference tasks.*

Remark 2 *There are other ways to characterize the law of a stochastic process. Arguably, the most popular approach is to use finite-dimensional time marginals (via Kolmogorov's extension theorem). However, this heavily depends on time-parametrization, whereas signature moments are parametrization invariant. Finally, other kernels for times series exist such as (Chan and Vasconcelos, 2005; Berndt and Clifford, 1994; Cuturi and Doucet, 2011; Moreno et al., 2003) as do methods coming from dynamic-warping; we refer to (Kiraly and Oberhauser, 2019, Section 1.4 and Section 5) for an attempt of an overview of this large literature. What distinguishes the signature kernel is that (i) it applies to general continuous time processes with discrete time arising as a special case—this ensures that it behaves well under high-frequency sampling (see Remark 3)—and (ii) it allows to use strong theoretical results from stochastic analysis which is essential for the proofs of our main theorems.*

What's Wrong With (Signature) Moments. Already the feature map from Example 1 runs into trouble if one replaces the compact subset $\mathcal{X}$ of $\mathbb{R}^d$ with a non-compact subset of $\mathbb{R}^d$: in this case, neither (A) or (B) hold in general; a well-known example is that a log-normal random variable is not characterized by its moments. Unfortunately this applies, a fortiori, to path-valued data and signature moments: even the laws of "trivial" stochastic processes that have straight lines as sample trajectories are in general not characterized by their signature moments; see Appendix B.5 for a concrete example.

Even worse, for vector-valued random variables a compact state space can be a reasonable assumption, but this is no longer the case for path-valued random variables since the space of paths is not even locally compact. A more reasonable alternative is to derive moment decay conditions that ensure that the sequence of moments still characterizes the law despite a non-compact support analogous to how a normally distributed (vector-valued) random variable is characterized by its moments. This was done in (Chevyrev and Lyons, 2016) for signature moments, see Example 2. However, these signature moment decay conditions firstly lead to strong assumptions that are hard to verify, in particular making them unsuitable for a non-parametric machine learning approach; secondly, such decay conditions exclude many important processes such as geometric Brownian motion.

Another related drawback of moments (classical and signature moments) is that of statistical robustness: already for $\mathbb{R}^d$-valued random variables, the mean $\mu \mapsto \mathbb{E}_{Y \sim \mu}[Y^{\otimes 1}] = \int y\mu(dy)$ is not robust to outliers, e.g. for mixtures of the form $\mu = (1 - \epsilon)\nu + \epsilon\delta_y$ the Dirac measure at y has too much influence, and the median would be a better "order $m = 1$" statistic; the same happens for higher moments $m > 1$ and such considerations motivated the development of "robust statistics", see (Hampel et al., 1986; Huber and Ronchetti, 2009).

1.1 Contribution

Our main contribution is to derive a "normalization" $\lambda(x)$ so that the *robust signature*

$$x \mapsto \Phi(x) = \left(\lambda(x)^m \int dx^{\otimes m}\right)_{m \geq 0} \tag{6}$$

(a) is a robust, characteristic and universal feature map on pathspace; more precisely

    (i) *universal* means that any continuous bounded function $f(x)$ of paths can be approximated by linear functionals of $\Phi(x)$,

    (ii) *characteristic* means that the map $\mu \mapsto \mathbb{E}_{X \sim \mu}[\Phi(X)]$, which maps the law of a stochastic process X to the expected value of $\Phi(X)$, is injective.

    (iii) *robust* means that the influence function (in the sense of Huber) of Φ is finite.

(b) gives rise to a characteristic and universal kernel $\mathrm{k}(x, y) = \langle \Phi(x), \Phi(y) \rangle$ for paths,

(c) the associated maximum mean discrepancy (MMD) is a metric for laws of stochastic processes that yields a topology comparable to the topology of weak convergence under natural assumptions.

The need for such kernel mean embeddings of laws of stochastic processes has been pointed out in the recent survey (Muandet et al., 2017, Section 6.2):

> *"Highly structured data can be taken into account using a kernel, but it is not clear how to employ kernel mean embedding for stochastic processes-i.e., distributions over infinite dimensional objects (...)."*

Our robust signature kernel and MMD provides an answer to this question; additionally, it provides the option to ignore the time-parameterization of stochastic processes, which can make the learning much more efficient by quotienting out invariance under a very large space of transformations. Some of our results remain interesting without the normalization λ, see Remark 4. Below we give a brief outline of our approach.

Robust Signatures. A recent observation is that universality and characteristicness are in duality and therefore often equivalent, see (Simon-Gabriel and Schölkopf, 2018).[2] Hence, it is sufficient to show universality, but this is more involved than in the finite-dimensional case due to the non-local compactness of pathspace. To address this, we use a convenient generalization of the Stone–Weierstrass theorem due to Giles (1971) together with a carefully constructed normalization map λ. We refer to the map Φ, as defined in (6), as *the robust signature* for two reasons: firstly, and most importantly, it is robust in the sense that no assumptions on the probability measure μ are needed to ensure characteristicness (except that the iterated integrals are well-defined). Secondly, the resulting signature moments $\mathbb{E}_{X \sim \mu}[\Phi(X)]$ that characterize the law μ of X are robust in the sense of statistical robustness to outliers mentioned above, (Hampel et al., 1986).

A MMD for Laws of Stochastic Processes. A natural distance between probability measures μ, ν on a topological space $\mathcal{X}$ is the "maximum mean discrepancy" (MMD)

$$d(\mu, \nu) = \sup_f |\mathbb{E}_{X \sim \mu}[f(X)] - \mathbb{E}_{Y \sim \nu}[f(Y)]| \ , \tag{7}$$

where the sup is taken over a sufficiently large set of functions from $\mathcal{X}$ to $\mathbb{R}$. To approximate (7) from finite samples of the laws μ (resp. ν), naive approaches are troublesome because of the supremum over a large space of functions. To address this, we follow (Berlinet and Thomas-Agnan, 2011; Gretton et al., 2012; Sriperumbudur et al., 2010) and kernelize the robust signature feature map Φ. That is, we define the normalized signature kernel $\mathrm{k}(x, y) := \langle \Phi(x), \Phi(y) \rangle$ and take the sup in (7) over functions in the unit ball of the reproducing kernel Hilbert space (henceforth RKHS) of k. This gives the identity

$$d_{\mathrm{k}}^2(\mu, \nu) = \mathbb{E}[\mathrm{k}(X, X')] - 2\mathbb{E}[\mathrm{k}(X, Y)] + \mathbb{E}[\mathrm{k}(Y, Y')] \ , \tag{8}$$

where X' and Y' denote independent copies of $X \sim \mu$ and $Y \sim \nu$. Efficient recursive algorithms for inner products of signatures have been developed, see (Kiraly and Oberhauser, 2019), and we can extend this to the robust signature Φ. This allows to evaluate $d_{\mathrm{k}}(\mu, \nu)$ from finite samples, even if the paths X, Y evolve in high-dimensional spaces (large d) or general topological spaces. Unfortunately, topologies induced by MMDs are typically hard to relate to standard topologies, see (Rachev, 1991). However, one of our main results is that d_{k} is a metric that induces a topology on probability measures strictly weaker than classical weak

---

2. The subtlety is that probability measures form a convex space but not a linear space, so one has to work with general distributions rather than probability measures for characteristicness.

convergence[3] (also called narrow convergence/convergence in law) and give conditions when these topologies are equal.

Application: Hypothesis Testing for Laws of Stochastic Processes. We apply our theoretical results to the problem of two-sample hypothesis testing for stochastic processes. A two-sample test for a stochastic processes $X = (X_t)_{t \in [0,T]}$, $Y = (Y_t)_{t \in [0,T]}$, tests the null-hypothesis

$$H_0 : P_X = P_Y \text{ against the alternative } H_1 : P_X \neq P_Y,$$

where $P_X := \mathbb{P} \circ X^{-1}$ and $P_Y := \mathbb{P} \circ Y^{-1}$ are the laws of X and Y. Our results about the characteristicness allow us to use the framework of Gretton et al. (2009, 2012) for this problem. The data we use is a popular archive of multi-variate time series data that contains heterogeneous datasets from various domains such as speech, handwriting, motion, etc. We benchmark against classical tests for multi-variate data and MMD tests using classical kernels.

1.2 Outline

Section 2 introduces the equivalence between universality and characteristicness of a feature map $\Phi : \mathcal{X} \to E$, as well as the strict topology that becomes useful when $\mathcal{X}$ is a general non-locally compact space. Section 3 focuses on the tensor algebra, $E := \mathbf{T}(V)$, as the natural feature space for moments and introduces the concept of a tensor normalization $\Lambda : \mathbf{T}_1(V) \to \mathbf{T}_1(V)$. As an application, we revisit Example 1 and show how such a normalization turns classical monomials on $\mathcal{X} = \mathbb{R}^d$ into a robust feature map for $\mathbb{R}^d$-valued random variables. Sections 4 and 5 introduce the robust signature map which extends this argument from $\mathbb{R}^d$-valued random variables to path-valued random variables. Section 6 introduces a bounded, continuous, universal and characteristic kernel $\mathrm{k}(x,y) = \langle \Phi(x), \Phi(y) \rangle$ for paths x, y and discusses the associated MMD on the space of laws of stochastic processes, the topology it induces, and its discretisations and computability. Section 7 contains as application a two-sample test between laws of stochastic processes. The Appendices contain technical details on rough paths, kernels, and the conducted experiments.

Remark 3 *Kiraly and Oberhauser (2019) introduced discretized signature kernels for sequence-valued data and provide efficient algorithms with a focus on supervised classification. We emphasize that, even for sequence-valued data, the question of characteristicness is non-trivial due to the lack of local compactness. This is even more the case for path-valued data, which we consider here. The move from sequence-valued to path-valued data is not only of theoretical interest: although real-world data sets often consist of sequences $(x(t_i))_{i=1,...,l}$ rather than paths $(x(t))_{t \in [0,T]}$, understanding the limiting behaviour as $\max_i |t_{i+1} - t_i| \to 0$ is crucial; e.g. it guarantees that no constants blow up in the high-frequency limit.*

---

3. This provides a natural example of a bounded, continuous, characteristic kernel which does not metrize weak convergence on a Polish space. The existence of such a kernel was raised as an open question in (Simon-Gabriel and Schölkopf, 2018). During the writing of this article, this question was additionally answered in (Simon-Gabriel et al., 2020) for even a locally compact Polish space. Additionally, the latter corrected an error in (Simon-Gabriel and Schölkopf, 2018) with regards to characterising kernels which metrize weak convergence.

Remark 4 *If one ignores the normalization, that is replaces $\lambda(x)$ by 1 in (6), then the signature features and kernel are in general not universal or characteristic; for example, the expected signature fails to characterize even some trivial stochastic processes, see Appendix B.5. However, their use might still be justified in certain situations (for example, if additional moment decay assumptions are known) and many of our results remain interesting in this case, e.g. the (non-robust) signature MMD can still be used for hypothesis testing, albeit under much stronger assumptions. Nevertheless, the computational cost of normalization is negligible and the robust signature (kernel) enjoys beyond universality and characteristicness also B-robustness which is in particular relevant for real-world data for the same reasons why robustness matters for vector-valued data; hence, in general we recommend the use of the robustified signature kernel resp. MMD.*

1.3 Notation

We collect some commonly used notation in the following table.

Symbol	Meaning	Page
	Spaces	
$\mathcal{X}$	topological space	
E	topological vector space (TVS) over $\mathbb{R}$	
E^*	algebraic dual of E, i.e., space of linear functionals $E \to \mathbb{R}$	
E'	topological dual of E, i.e., subspace $E' \subset E^*$ of *continuous* functionals	
H	Hilbert space over $\mathbb{R}$	
V	Banach space over $\mathbb{R}$	
$\overline{V}$	$\overline{V} = V \oplus \mathbb{R}$	17
	Paths and sequences	
$\mathcal{C}^1$	the subset of $\cup_{T>0} C([0,T], V)$ of absolutely continuous paths	14
ρ	pseudo-metric on $\mathcal{C}^1$	15
$\overline{\rho}$	metric on $\mathcal{C}^1(\overline{V})$	17
$\mathcal{P}^1$	tree-like equivalence classes of absolutely continuous paths in V	16
$x \sim_t y$	tree-like equivalence relation between $x, y \in \cup_{T>0} C([0,T], V)$	33
π	partition of $[0,T]$, i.e., a collection of points $0 \leq t_1 < \cdots < t_l \leq T$	
x^π	the sequence $x^\pi := (x(t_i))_{i=1,\dots,l}$ given by sampling x along π	
	Signatures and normalization	
$\int dx^{\otimes m}$	shorthand for $\int_{0 \leq t_1 \leq \cdots \leq t_m \leq T} dx(t_1) \otimes \cdots \otimes dx(t_m)$	15
S	signature map, $\mathrm{S}(x) = (\int dx^{\otimes m})_{m \geq 0}$	15
$\mathbf{T}(V)$	square summable elements in the tensor algebra $\prod_{m \geq 0} V^{\otimes m}$	11
$\mathbf{T}_1(V)$	subset of $\mathbf{T}(V)$ with zero-th component equal 1	11
Λ	a tensor normalization	11

2. Learning in Non-locally Compact Spaces

As discussed in the Introduction, two classical problems of learning in a topological space $\mathcal{X}$ are: (A) make inference about a function $f \in \mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$, and (B) make inference about a probability measure μ on $\mathcal{X}$. A standard approach is to map $\mathcal{X}$ into a (typically infinite or high dimensional) linear space E and address the learning problem there by using linear methods. All vector spaces in the sequel are over $\mathbb{R}$.

Definition 5 *Let $\mathcal{X}$ be a topological space. For a topological vector space (TVS) E, we call any map $\Phi : \mathcal{X} \to E$ a feature map for which E is the feature space. We denote by E' and E^* the topological and algebraic duals of E respectively.*

Universal and Characteristic Features.   To address Point (A), we require that E' is large enough to approximate the elements of the function class $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$, i.e. for every $f \in \mathcal{F}$, there exists $\ell \in E'$ such that $f(\cdot) \approx \langle \ell, \Phi(\cdot) \rangle$ as real-valued functions on $\mathcal{X}$. To address Point (B), we require that the feature map is non-linear enough to distinguish measures μ on $\mathcal{X}$, i.e. the map

$$\mu \mapsto \mu(\Phi) := \int_{\mathcal{X}} \Phi(x) \mu(dx) \in E$$

is injective. Since E can be infinite dimensional, it is more convenient to write this map as

$$\mu \mapsto \left(\ell \mapsto \int_{\mathcal{X}} \langle \ell, \Phi(x) \rangle \mu(dx) \right) \in \left(E' \right)^* .$$

More generally, we can replace the integral by any distribution $D \in \mathcal{F}'$, i.e. require that

$$D \mapsto (\ell \mapsto D(x \mapsto \langle \ell, \Phi(x) \rangle))$$

is injective. The definition of universality and characteristicness of Φ makes this precise.

Definition 6 *Fix $\mathcal{X}$ and a TVS $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$. Consider a feature map $\Phi : \mathcal{X} \to E$ such that $\langle \ell, \Phi(\cdot) \rangle \in \mathcal{F}$ for all $\ell \in E'$. We say that Φ is*

   (a) *universal to $\mathcal{F}$ if the following map has a dense image:*

$$\imath : E' \to \mathcal{F}, \quad \ell \mapsto \langle \ell, \Phi(\cdot) \rangle . \tag{9}$$

   (b) *characteristic to a subset $\mathcal{P} \subset \mathcal{F}'$ if the following map is injective:*

$$\kappa : \mathcal{P} \to (E')^*, \quad D \mapsto [\ell \mapsto D(\langle \ell, \Phi(\cdot) \rangle)] .$$

Universality directly addresses Point (A); characteristicness addresses Point (B) in a much more general sense since the dual $\mathcal{F}'$ is larger than the set of probability measures on $\mathcal{X}$. This generalization allows to work on the linear space $\mathcal{F}'$ (instead of a convex subset) and allows for the following simple statement of duality.

Theorem 7 *Suppose that $\mathcal{F}$ is a locally convex TVS. A feature map Φ is universal to $\mathcal{F}$ iff Φ is characteristic to $\mathcal{F}'$.*

The proof of Theorem 7 employs the Hahn–Banach theorem and is identical to that of a corresponding result for kernels (Simon-Gabriel and Schölkopf, 2018, Thm. 6). We will see that this result is useful since universality is often much easier to show than characteristicness.

Example 3 (Example 1 revisited) *Here, $\mathcal{X} \subset \mathbb{R}^d$ compact, $E := \prod_{m \geq 0} \left(\mathbb{R}^d \right)^{\otimes m}$ and $\Phi(x) := \left(\frac{x^{\otimes m}}{m!} \right)_{m \geq 0}$. The image $\iota(E')$ is the space of polynomials on $\mathcal{X}$. By Stone–Weierstrass, it follows that Φ is universal to $\mathcal{F} := C(\mathcal{X}, \mathbb{R})$ equipped with the uniform topology. By Theorem 7, Φ is characteristic to $\mathcal{F}'$ which is the space of regular Borel measure on $\mathcal{X}$. In particular, this implies the well-known result that the law of a bounded random variable is determined by its moments.*

2.1 The Strict Topology

To extend Example 3 from a compact subset $\mathcal{X}$ of $\mathbb{R}^d$ to a general non-compact data domain $\mathcal{X}$, we need to

(a) find a map $\Phi : \mathcal{X} \to E$ that behaves like "monomials" so that linear functionals and consequently linear functionals $\langle \ell, \Phi(x) \rangle$ behave like "polynomials" in x; put more abstractly, $\mathcal{F}_0 := \{ x \mapsto \langle \ell, \Phi(x) \rangle : \ell \in E' \}$ should form a point-separating algebra,

(b) find a TVS $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ such that $\mathcal{F}'$ includes the probability measures on $\mathcal{X}$,

(c) have a Stone–Weierstrass-type result that shows that $\mathcal{F}_0$ is dense in $\mathcal{F}$.

Point (a) must take into account the concrete choice of $\mathcal{X}$. Unfortunately, already for the arguably simple case $\mathcal{X} = \mathbb{R}^d$ we know that the standard monomial map from Examples 1 and 3 does not work since there exist unbounded random variables that have different laws but the same moment sequence, see Appendix B.5. Further, for non-locally compact spaces $\mathcal{X}$ (like the space of paths) it is often easy to find a function class $\mathcal{F}$ such that exactly one of (b) and (c) holds, but not both; typically the dual space $\mathcal{F}'$ is either too small, or Stone–Weierstrass-type results are not known or involve conditions that are hard to verify.

Below we recall a result of Giles (1971) which provides a surprisingly elegant and general roadmap on what is needed to addresses all points (a), (b), and (c).

Definition 8 *Let $\mathcal{X}$ be a topological space. We say that a function $\psi : \mathcal{X} \to \mathbb{R}$ vanishes at infinity if for all $\epsilon > 0$ there exists a compact set $K \subset \mathcal{X}$ such that $\sup_{x \in \mathcal{X} \setminus K} |\psi(x)| < \epsilon$. Denote with $B_0(\mathcal{X}, \mathbb{R})$ the set of functions that vanish at infinity. The strict topology on $C_b(\mathcal{X}, \mathbb{R})$ is the topology generated by the seminorms*

$$p_\psi(f) = \sup_{x \in \mathcal{X}} |f(x)\psi(x)| , \quad \psi \in B_0(\mathcal{X}, \mathbb{R}) .$$

Theorem 9 (Giles (1971)) *Let $\mathcal{X}$ be a metrizable topological space.*

(i) The strict topology on $C_b(\mathcal{X}, \mathbb{R})$ is weaker than the uniform topology and stronger than the topology of uniform convergence on compact sets.

(ii) If $\mathcal{F}_0$ is a subalgebra of $C_b(\mathcal{X}, \mathbb{R})$ such that

   (a) $\forall x \neq y \in \mathcal{X}$ there exists $f \in \mathcal{F}_0$ such that $f(x) \neq f(y)$, and

   (b) $\forall x \in \mathcal{X}$ there exists $f \in \mathcal{F}_0$ such that $f(x) \neq 0$,

   then $\mathcal{F}_0$ is dense in $C_b(\mathcal{X}, \mathbb{R})$ under the strict topology.

(iii) The topological dual of $C_b(\mathcal{X}, \mathbb{R})$ equipped with the strict topology is the space of finite regular Borel measures on $\mathcal{X}$.

Proof Point (i) follows from the definition. Point (ii) is (Giles, 1971, Thm. 3.1); in fact, the result holds more generally for any topological space and one only needs that point-separating and non-vanishing functions are in the closure of $\mathcal{F}_0$. Point (iii) is (Giles, 1971, Thm. 4.6). ∎

The specific construction of Φ must depend on the domain $\mathcal{X}$. However, often one has a natural candidate for monomials on $\mathcal{X}$. For example, for $\mathcal{X} = \mathbb{R}^d$ one can use the standard monomials as used in Examples 1 and 3; for $\mathcal{X}$ the space of paths, stochastic analysis suggests that the iterated integrals $\int dx^{\otimes m}$ are the "natural monomials" of degree m. It remains to have a method that generically turns such monomials into "bounded monomials" to which Theorem 9 applies. In Section 3 we introduce a generic normalization that accomplishes this. In Section 4 we apply this construction to the space of paths to derive our robust signature features.

3. Robustification by Normalization

Already the standard monomials in $\mathbb{R}^d$, Example 1, are unbounded on non-compact domains $\mathcal{X}$, hence linear functionals of them (polynomials) are not bounded and thus Theorem 9 does not apply. On the other hand, polynomials have many nice properties: they are closed under multiplication and they separate points. In this section we show that there exists a generic "normalization" that makes monomials bounded while preserving their algebra and point-separating property. Moreover, this normalization also turns them into robust statistics in the sense of Huber and Ronchetti (2009). Even for the case $\mathcal{X} = \mathbb{R}^d$ this leads to a novel moment sequence that characterize the law of any $\mathbb{R}^d$-valued random variable, Section 3.4. More importantly for our purposes, this method also applies to signature features on paths.

3.1 The Tensor Algebra

Throughout the rest of this article let us denote with V a Banach space. Recall that $V^{\otimes m}$ denotes the space of tensors of degree m. A *sequences of tensors of increasing degree* is an element of the space $\prod_{m \geq 0} V^{\otimes m}$. We will denote elements $\mathbf{t} \in \prod_{m \geq 0} V^{\otimes m}$ by

$$\mathbf{t} = (\mathbf{t}^m)_{m \geq 0} \in \prod_{m \geq 0} V^{\otimes m}$$

where $\mathbf{t}^m \in V^{\otimes m}$. The space $\prod_{m \geq 0} V^{\otimes m}$ becomes a linear space by defining

$$\mathbf{s} + \mathbf{t} := \left(\mathbf{s}^0 + \mathbf{t}^0, \mathbf{s}^1 + \mathbf{t}^1, \ldots\right) \text{ for } \mathbf{s} = (\mathbf{s}^m)_{m \geq 0}, \mathbf{t} = (\mathbf{t}^m)_{m \geq 0} \in \prod_{m \geq 0} V^{\otimes m}.$$

This space is the natural state space for monomials; for general background on tensors we refer to Appendix A. Of particular relevance is the subspace $\mathbf{T}(V) \subset \prod_{m \geq 0} V^{\otimes m}$ of elements that have a finite norm.

Definition 10 *Let V be a Banach space.[4] We denote by $\mathbf{T}(V)$ the Banach space*

$$\mathbf{T}(V) := \Big\{ \mathbf{t} \in \prod_{m \geq 0} V^{\otimes m} \: : \: \|\mathbf{t}\| := \Big(\sum_{m \geq 0} \|\mathbf{t}^m\|_{V^{\otimes m}}^2 \Big)^{1/2} < \infty \Big\} \: .$$

Define further the subset $\mathbf{T}_1(V) := \{ \mathbf{t} \in \mathbf{T}(V) \: : \: \mathbf{t}^0 = 1 \}$.

Remark 11 *If $V = H$ is a Hilbert space, we will always equip $H^{\otimes m}$ with the canonical inner product (which induces an admissible system of norms) given on elementary tensors by*

$$\langle x_1 \otimes \ldots \otimes x_m, y_1 \otimes \ldots \otimes y_m \rangle_{H^{\otimes m}} = \prod_{j=1}^{m} \langle x_j, y_j \rangle_H \: ,$$

and extended by linearity to $H^{\otimes m}$. Consequently, $\mathbf{T}(H)$ becomes a Hilbert space with inner product $\langle \mathbf{s}, \mathbf{t} \rangle_{\mathbf{T}(H)} = \sum_{m \geq 0} \langle \mathbf{s}^m, \mathbf{t}^m \rangle_{H^{\otimes m}}$.

3.2 Normalization

Naive approaches to make monomials bounded such as dividing them by their norm destroys that they form an algebra and that they are point-separating. The idea of a tensor normalization, is that these "nice" properties of monomials can be preserved while making them bounded.

Definition 12 *A* tensor normalization *is a continuous injective map of the form*

$$\Lambda : \mathbf{T}_1(V) \to \{ \mathbf{t} \in \mathbf{T}_1(V) : \|\mathbf{t}\| \leq R \} \: , \qquad \Lambda : \mathbf{t} \mapsto \delta_{\lambda(\mathbf{t})} \mathbf{t} \: ,$$

where $R > 0$ is a constant, $\lambda : \mathbf{T}_1(V) \to (0, \infty)$ is a function, and, for $c \in \mathbb{R}$, δ_c is the dilation map $\delta_c(\mathbf{t}) = (\mathbf{t}^0, c\mathbf{t}^1, c^2\mathbf{t}^2, \ldots)$.

The existence of tensor normalizations is not trivial. In the rest of this section, we give a general method to construct such maps and determine their regularity properties. In Sections 6.4 and 7.3 we show that tensor normalization can be carried out in a kernel learning setting with minor computational cost.

Lemma 13 *Suppose that $\lambda \geq 0$ and $\mathbf{t} \in \mathbf{T}(V)$ such that $\delta_\lambda \mathbf{t} \in \mathbf{T}(V)$. Then $\|\delta_\lambda \mathbf{t} - \mathbf{t}\|^2 \leq |\|\delta_\lambda \mathbf{t}\|^2 - \|\mathbf{t}\|^2|$.*

Proof We may suppose without loss of generality that $\lambda \geq 1$ (otherwise we swap $\mathbf{t}$ with $\delta_\lambda \mathbf{t}$ and λ with λ^{-1}). Taking derivatives in $\bar{\lambda}$ at some $\bar{\lambda} \in [1, \lambda)$,

$$\frac{d}{d\bar{\lambda}} \|\delta_{\bar{\lambda}} \mathbf{t} - \mathbf{t}\|^2 = \sum_{m=1}^{\infty} 2m(\bar{\lambda}^{2m-1} - \bar{\lambda}^{m-1}) \|\mathbf{t}^m\|_{V^{\otimes m}}^2 \leq \sum_{m=1}^{\infty} 2m\bar{\lambda}^{2m-1} \|\mathbf{t}^m\|_{V^{\otimes m}}^2 = \frac{d}{d\bar{\lambda}} \|\delta_{\bar{\lambda}} \mathbf{t}\|^2 \: ,$$

---

4. We implicitly assume that $V^{\otimes m}$ for all $m \geq 2$ are Banach spaces formed by the completion with respect to some admissible system of tensor norms $\| \cdot \|_{V^{\otimes m}}$, see (Lyons and Qian, 2002, Sec. 3.1).

where all sums are convergent by the assumption that $\sum_{m=1}^{\infty} \lambda^{2m} \|\mathbf{t}^m\|^2 < \infty$. Since both derivatives are positive and the first is bounded above by the second, the conclusion follows. ∎

Proposition 14 *Let $\psi : [1, \infty) \to [1, \infty)$ with $\psi(1) = 1$. For $\mathbf{t} \in \mathbf{T}_1(V)$, let $\lambda(\mathbf{t}) \geq 0$ denote the unique non-negative number such that $\|\delta_{\lambda(\mathbf{t})}\mathbf{t}\|^2 = \psi(\|\mathbf{t}\|)$. Define $\Lambda : \mathbf{T}_1(V) \to \mathbf{T}_1(V)$, $\Lambda(\mathbf{t}) = \delta_{\lambda(\mathbf{t})}\mathbf{t}$. Denote further $\|\psi\|_\infty = \sup_{x \geq 1} \psi(x)$.*

  (i) It holds that Λ takes values in the set $\{\mathbf{t} \in \mathbf{T}_1(V) : \|\mathbf{t}\| \leq \sqrt{\|\psi\|_\infty}\}$.

  (ii) If ψ is injective, then so is Λ.

  (iii) Suppose that $\sup_{x \geq 1} \psi(x)/x^2 \leq 1$, $\|\psi\|_\infty < \infty$, and that ψ is K-Lipschitz for some $K > 0$. Then

$$\|\Lambda(\mathbf{s}) - \Lambda(\mathbf{t})\| \leq (1 + K^{1/2} + \sqrt{2}\|\psi\|_\infty^{1/2})(\|\mathbf{s} - \mathbf{t}\|^{1/2} \vee \|\mathbf{s} - \mathbf{t}\|) \ .$$

Proof (i) is clear by construction. For (ii) suppose that ψ is injective and that $\Lambda(\mathbf{t}) = \Lambda(\mathbf{s})$. If $\mathbf{t} = \mathbf{1} := (1, 0, 0 \ldots)$, then evidently $\mathbf{s} = \mathbf{t}$. We thus suppose $\mathbf{t} \neq \mathbf{1}$. By definition of Λ, it holds that $\mathbf{s} = \delta_\lambda \mathbf{t}$ for some $\lambda \geq 0$. On the other hand, we also have $\psi(\|\mathbf{t}\|) = \psi(\|\mathbf{s}\|)$. Since ψ is injective and $\lambda \mapsto \|\delta_\lambda \mathbf{t}\|$ is strictly increasing by the assumption that $\mathbf{t} = \mathbf{1}$, it follows that $\mathbf{s} = \mathbf{t}$, which proves (ii). For (iii), note that $\psi(x) \leq x^2$ implies $\lambda(\mathbf{t}) \leq 1$, and thus $\delta_{\lambda(\mathbf{t})}$ is 1-Lipschitz on $\mathbf{T}_1(V)$. Hence

$$\begin{aligned}
\|\delta_{\lambda(\mathbf{s})}\mathbf{s} - \delta_{\lambda(\mathbf{t})}\mathbf{t}\| &\leq \|\delta_{\lambda(\mathbf{t})}\mathbf{s} - \delta_{\lambda(\mathbf{t})}\mathbf{t}\| + \|\delta_{\lambda(\mathbf{s})}\mathbf{s} - \delta_{\lambda(\mathbf{t})}\mathbf{s}\| \\
&\leq \|\mathbf{s} - \mathbf{t}\| + |\|\delta_{\lambda(\mathbf{s})}\mathbf{s}\|^2 - \|\delta_{\lambda(\mathbf{t})}\mathbf{s}\|^2|^{1/2} \\
&\leq \|\mathbf{s} - \mathbf{t}\| + |\psi(\|\mathbf{s}\|) - \psi(\|\mathbf{t}\|)|^{1/2} + |\|\delta_{\lambda(\mathbf{t})}\mathbf{t}\|^2 - \|\delta_{\lambda(\mathbf{t})}\mathbf{s}\|^2|^{1/2} \\
&\leq \|\mathbf{s} - \mathbf{t}\| + K^{1/2}\|\mathbf{s} - \mathbf{t}\|^{1/2} + \sqrt{2}\|\psi\|_\infty^{1/2}\|\mathbf{s} - \mathbf{t}\|^{1/2} \ ,
\end{aligned}$$

where we used in the second line Lemma 13 and in the fourth line that ψ is K-Lipschitz and without loss of generality that $\lambda(\mathbf{t}) \leq \lambda(\mathbf{s})$ (hence $\|\delta_{\lambda(\mathbf{t})}\mathbf{s}\| \leq \|\delta_{\lambda(\mathbf{s})}\mathbf{s}\| \leq \|\psi\|_\infty$). ∎

Corollary 15 *Let $\psi : [1, \infty) \to [1, \infty)$ be injective satisfying $\psi(1) = 1$ and the conditions of Proposition 14(iii). Then Λ constructed in Proposition 14 is a tensor normalization.*

Example 4 *The tensor normalization Λ we use in our experiments is built as in Corollary 15 from the function $\psi : [1, \infty) \to [1, \infty)$ defined by*

$$\psi(\sqrt{x}) = \begin{cases} x & \text{if } x \leq C \ , \\ C + C^{1+a}(C^{-a} - x^{-a})/a & \text{if } x > C \ , \end{cases}$$

where $a > 0, C \geq 1$ are constants. Note that ψ satisfies all the conditions of Corollary 15; in particular, ψ is $2C$-Lipschitz and bounded above by $C(1 + \frac{1}{a})$. Figure 1 shows the graph of $\psi(\sqrt{x})$ with $a = 1$ and $C = 4$; note that $\Lambda(\mathbf{t}) = \mathbf{t}$ if $\|\mathbf{t}\| \leq 4$.
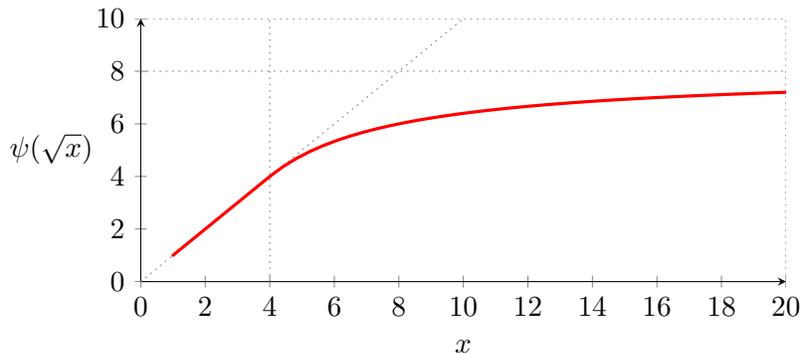
Figure 1: The graph of $\psi(\sqrt{x})$ from Example 4 with $a = 1$ and $C = 4$.

3.3 B-robust Statistics

In addition to allowing us to applying Theorem 9, tensor normalizations guarantee that the statistics which are linear functionals $\langle \ell, \Phi(X) \rangle$ of a normalized feature map $\Phi = \Lambda \circ \varphi$ are robust to outliers in a precise statistical sense.

Definition 16 (Hampel et al. (1986)) *Let* $\mathcal{P}(\mathcal{X})$ *denote the space of Borel probability measures on* $\mathcal{X}$. *For a map* $T : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ *define the* influence function *of* T *at* $\mu \in \mathcal{P}(\mathcal{X})$ *by*

$$\mathrm{IF}(x; T, \mu) := \lim_{\epsilon \searrow 0} \frac{T(\epsilon \delta_x + (1 - \epsilon)\mu) - T(\mu)}{\epsilon}$$

for $x \in \mathcal{X}$ *for which this limit exists.* T *is called* B-robust *at* μ *if* $\sup_{x \in \mathcal{X}} |\mathrm{IF}(x; T, \mu)| < \infty$.

Example 5 *Suppose* $f : \mathcal{X} \to \mathbb{R}$ *is measurable and* $T(\mu) := \mu(f) = \int_{\mathcal{X}} f(x)\mu(dx)$ *if* f *is* μ *integrable (and arbitrary otherwise). Then, provided* f *is* μ-integrable, $\mathrm{IF}(x; T, \mu) = f(x) - \mu(f)$, *and therefore* T *is* B-robust at μ *if and only if* f *is bounded.*

The previous example shows that if $\mathcal{X} \subset \mathbb{R}$ and $T : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ is the m-th moment $T(\mu) = \mathbb{E}_{X \sim \mu}[X^m]$ for some $m \geq 1$, then T is B-robust (at any μ with finite m-th moment) if and only if $\mathcal{X}$ is bounded. The same example, however, shows the following.

Proposition 17 *Consider* Λ *a tensor normalization,* $\varphi : \mathcal{X} \to \mathbf{T}_1(V)$ *measurable, and* $\ell \in \mathbf{T}(V)'$. *Define* $\Phi = \Lambda \circ \varphi$ *and* $f(x) = \langle \ell, \Phi(x) \rangle$. *Then* $\mu \mapsto \mu(f)$ *is* B-robust at any μ.

The significance of this result is that linear functionals of the feature maps 'Φ' appearing in our results below (Proposition 18, Theorems 21 and 26, etc.) are automatically B-robust statistics. The focus of this article is to characterize probability measures on pathspace, but the above observation that B-robustness is intimately linked to moment normalization and characteristicness is already interesting for $\mathbb{R}^d$-valued random variables.

3.4 Example: Robust Moments for Vector-valued Data

We now demonstrate the usefulness of the tensor normalization by turning the standard monomial feature map from Example 3 into a robust, characteristic, and universal feature map on the whole of $\mathbb{R}^d$. We emphasise that our argument does not rely on local compactness and thus paves the way to apply the same reasoning to pathspace.

Proposition 18 *Consider the map*

$$\varphi : \mathbb{R}^d \to \mathbf{T}_1(\mathbb{R}^d); , \quad x \mapsto \Big(\frac{x^{\otimes m}}{m!}\Big)_{m=0}^{\infty} . \tag{10}$$

Let $\Lambda : \mathbf{T}_1(\mathbb{R}^d) \to \mathbf{T}_1(\mathbb{R}^d)$ be a tensor normalization. Then the map

$$\Phi : \mathbb{R}^d \to \mathbf{T}_1(\mathbb{R}^d), \quad \Phi = \Lambda \circ \varphi$$

(i) *is a continuous injection from $\mathbb{R}^d$ into a bounded subset of $\mathbf{T}_1(\mathbb{R}^d)$,*

(ii) *is universal to $\mathcal{F} = C_b(\mathbb{R}^d, \mathbb{R})$ equipped with the strict topology,*

(iii) *is characteristic to $\mathcal{F}'$, the set of finite, signed Borel measures on $\mathbb{R}^d$.*

Moreover, for any $\ell \in \mathbf{T}(V)'$ the map $\mu \to \langle \ell, \mathbb{E}_{X \sim \mu}[\Phi(X)] \rangle$ is B-robust.

Proof (i) follows from the definition of the tensor normalization Λ and the fact that $\varphi : \mathbb{R}^d \to \mathbf{T}_1(\mathbb{R}^d)$ is a continuous injection. To show (ii), we claim that the family of functions $\mathcal{F}_0 = \{x \mapsto \langle \ell, \Phi(x) \rangle, \ell \in (\mathbb{R}^m)', \ m \geq 0\} \subset C_b(\mathbb{R}^d, \mathbb{R})$ satisfies the conditions of Theorem 9(ii). Indeed, for $m, n \geq 0$ and $\ell \in (\mathbb{R}^d)^{\otimes m}, \tilde{\ell} \in (\mathbb{R}^d)^{\otimes n}$ we have $\langle \ell, \Phi(x) \rangle \langle \tilde{\ell}, \Phi(x) \rangle = \lambda(\varphi(x))^{m+n} \langle \ell, \frac{x^{\otimes m}}{m!} \rangle \langle \tilde{\ell}, \frac{x^{\otimes n}}{n!} \rangle = \binom{m+n}{m} \langle \ell \otimes \tilde{\ell}, \Phi(x) \rangle$, which shows that $\mathcal{F}_0$ is a subalgebra of $C_b(\mathbb{R}^d, \mathbb{R})$. The family $\mathcal{F}_0$ separates points since Φ is injective. Finally, note that $\Phi(x) \neq 0$ for all $x \in \mathbb{R}^d$. It follows by Theorem 9(ii) that Φ is universal to $C_b(\mathbb{R}^d, \mathbb{R})$ equipped with the strict topology which shows (ii); (iii) now follows by Theorem 9(iii). ∎

An informal point of view of Proposition 18 is to recall the difference between the moment generating function and the characteristic function of $\mathbb{R}^d$-valued random variable X. The former can fail to characterize laws of unbounded random variables, whereas the latter characterizes the law of any random variable. The characteristic function accomplishes this by using a Fourier transform that spins faster in a circle in the complex plane for large value of X. The above proposition shows that the same can be accomplished by "smearing out" large realizations of X into a ball in $\mathbf{T}(\mathbb{R}^d)$.

4. Robust Features for Smooth Paths

In this section, we apply the robustification introduced in Section 3 on pathspace. We first focus on the case when $\mathcal{X} = \mathcal{C}^1$ denotes the subset of $\cup_{T>0} C([0, T], V)$ of absolutely continuous paths. The case for more irregular paths is covered in Section 5.

4.1 Iterated Integrals: Monomials of Paths

We consider first absolutely continuous paths parameterized over any finite interval $[0, T]$ whose starting point is the origin:

$$\mathcal{C}^1(V) := \bigcup_{T>0} \Big\{ x \in C([0, T], V) : \exists \dot{x} \in L^1([0, T], V) \, , \, x = \int_0^{\cdot} \dot{x}(t) dt \Big\} ,$$

$$\|x\|_{1-\text{var};[0,T]} := \int_0^T \|\dot{x}(t)\| dt = \sup_{\substack{n \geq 1 \\ 0 \leq t_1 \leq \cdots \leq t_n \leq T}} \sum_{i=1}^{n-1} \|x(t_{i+1}) - x(t_i)\| .$$

When the space V is clear from the context we just write $\mathcal{C}^1$ instead of $\mathcal{C}^1(V)$. When we speak of an element $x \in \mathcal{C}^1$, we let $[0, T_x]$ denote the corresponding time interval. Due to the arbitrary time interval $[0, T]$, $\mathcal{C}^1$ is not a vector space (since two paths on different time intervals can not be added). To define a topology on $\mathcal{C}^1$, consider for $x, y \in \mathcal{C}^1$ the quantity[5]

$$\rho(x, y) = \inf_\tau \|x - y \circ \tau\|_{1-\text{var};[0,T_x]} , \tag{11}$$

where the infimum is taken over all increasing bijections $\tau : [0, T_x] \to [0, T_y]$; we call such a τ a reparameterization. Note that ρ satisfies all the properties of a metric except that $\rho(x, y) = 0$ does in general not imply that $x = y$ (i.e., ρ is a pseudo-metric). We equip $\mathcal{C}^1$ with the (non-Hausdorff) topology induced by ρ. The $\mathcal{C}^1$ regularity of the paths allows us to use Riemann–Stieltjes integration to define (see also (4))

$$\int_0^t dx^{\otimes m} := \int_{0 \le t_1 \le \cdots \le t_m \le t} dx(t_1) \otimes \cdots \otimes dx(t_m) \in V^{\otimes m} \quad \text{for } t \in [0, T_x] . \tag{12}$$

We will often use the shorthand $\int dx^{\otimes m} = \int_0^{T_x} dx^{\otimes m}$. A guiding principle that plays a prominent role in stochastic analysis is that the so-called *signature map*

$$\mathrm{S} : \mathcal{C}^1 \to \mathbf{T}_1(V) , \quad x \mapsto \left(1, \int dx, \int dx^{\otimes 2}, \int dx^{\otimes 3}, \dots\right) , \tag{13}$$

is the natural generalization of the monomial feature map from Example 1 to pathspace.

Example 6 *Let $V = \mathbb{R}^d$. Consider the linear path $x : [0, 1] \to \mathbb{R}^d$, $x(t) = tv$, for a fixed vector $v \in \mathbb{R}^d$. A simple calculation shows $\int dx^{\otimes m} = \frac{v^{\otimes m}}{m!}$. Hence, S is a generalization of the monomial feature map on $\mathbb{R}^d$ from Example 1.*

An algebraic reason why (12) can be seen as sequence of monomials on pathspace $\mathcal{C}^1$ is that their linear span is closed under multiplication. The so-called shuffle product makes this precise: considering for simplicity $V = \mathbb{R}^d$, we have for $\mathbf{i} = (i_1, \dots, i_m) \in \{1, \dots, d\}^m$, $\mathbf{j} = (j_1, \dots, j_n) \in \{1, \dots, d\}^n$

$$\int dx_{\mathbf{i}}^{\otimes m} \int dx_{\mathbf{j}}^{\otimes n} = \sum_{\mathbf{k}} \int dx_{\mathbf{k}}^{\otimes(m+n)} , \tag{14}$$

where the sum is taken over all $\mathbf{k} = (k_1, \dots, k_{m+n})$ that are shuffles[6] of $\mathbf{i}$ and $\mathbf{j}$, see (Lyons et al., 2007, Thm. 2.15). The signature map furthermore identifies paths up to tree-like equivalence.

Theorem 19 (Boedihardjo et al. (2016)) *The map $\mathrm{S} : \mathcal{C}^1 \to \mathbf{T}_1(V)$ is injective up to tree-like equivalence: for $x, y \in \mathcal{C}^1$, $\mathrm{S}(x) = \mathrm{S}(y)$ if and only if $x \sim_t y$.*

The tree-like equivalence relation $x \sim_t y$ is a natural generalization of the reparameterization relation on pathspace, see Footnote 1 and Appendix B.1. In particular, and perhaps most importantly, two paths x, y are tree-like equivalent if they are reparameterizations, $x(t) = y(\tau(t))$ for all t (or more generally if $\rho(x, y) = 0$). We show in Section 4.3, that, if desired, it is easy to remove the tree-like equivalence relation in Theorem 19. Nevertheless, the identification of tree-like equivalent paths (read: reparameterized paths) can be a powerful dimensionality reduction.

---

5. Readers will notice the similarity with dynamic time warping and Fréchet distance.
6. A shuffle $\mathbf{k} = (k_1, \dots, k_{m+n})$ of $\mathbf{i}$ and $\mathbf{j}$ is a permutations of $(i_1, \dots, i_m, j_1, \dots, j_n)$ subject to the condition that the order of elements in $\mathbf{i}$ and $\mathbf{j}$ is preserved.

4.2 Features for Unparameterized Paths

The empirical success of dynamic time warping and Fréchet distances show that for path-valued data it can beneficial to ignore the time-parameterization. Motivated by this we identify paths that are reparameterizations. Hence, instead of the space of paths $\mathcal{C}^1$ we first derive robust features for the space of equivalence classes of paths $\mathcal{P}^1 := \mathcal{C}^1 / \sim_t$.

Definition 20 *We define the space of unparameterized absolutely continuous paths $\mathcal{P}^1$ as the set of equivalence classes $\mathcal{C}^1 / \sim_t$. We equip $\mathcal{P}^1$ with the initial topology[7] induced by the signature map embedding $\mathrm{S} : \mathcal{C}^1 \to \mathbf{T}_1(V)$.*

In the above definition, recall that given a set $\mathcal{Y}$ and a topological space $\mathcal{X}$, the *initial topology* on $\mathcal{Y}$ induced by a function $f : \mathcal{Y} \to \mathcal{X}$ is the weakest topology on $\mathcal{Y}$ such that f is continuous; remark that we can view S as a function $\mathcal{P}^1 \to \mathbf{T}_1(V)$ due to Theorem 19.

In complete analogy to our toy example of the monomial feature map for $\mathbb{R}^d$ in Proposition 18, we apply a tensor normalization to the signature map S to get a robust, universal, and characteristic feature map.

Theorem 21 *Let $\Lambda : \mathbf{T}_1(V) \to \mathbf{T}_1(V)$ be a tensor normalization. The robust signature*

$$\Phi : \mathcal{P}^1 \to \mathbf{T}_1(V), \quad \Phi = \Lambda \circ \mathrm{S}$$

(i) is a continuous injection from $\mathcal{P}^1$ into a bounded subset of $\mathbf{T}_1(V)$,

(ii) is universal to $\mathcal{F} := C_b(\mathcal{P}^1, \mathbb{R})$ equipped with the strict topology,

(iii) is characteristic to the space of finite regular Borel measures on $\mathcal{P}^1$.

Moreover, $\forall \ell \in \mathbf{T}(V)'$ the map $\mu \to \langle \ell, \mathbb{E}_{X \sim \mu}[\Phi(X)] \rangle$ is B-robust.

Proof (i) follows from the definition of tensor normalization and the fact that $\mathrm{S} : \mathcal{P}^1 \to \mathbf{T}_1(V)$ is a homeomorphism onto its image. For (ii), define $L := \bigoplus_{m \geq 0} (V^{\otimes m})'$, which we identify with a dense subspace of $\mathbf{T}(V)'$ via $\ell(\mathbf{t}) = \sum_{m \geq 0} \langle \ell^m, \mathbf{t}^m \rangle$. Define further $\mathcal{F}_0 := \{\ell \circ \Phi : \ell \in L\} \subset \mathcal{F}$. By point (i), $\mathcal{F}_0$ separates the points $\mathcal{P}^1$. Furthermore, S takes values in the group-like elements of $\mathbf{T}(V)$ (Cass et al., 2016, Cor. 3.9) which implies $\mathcal{F}_0$ is closed under multiplication (in the case $V = \mathbb{R}^d$, this is equivalent to the shuffle product (14)). It follows that $\mathcal{F}_0$ satisfies the assumptions of Theorem 9(ii). Hence $\mathcal{F}_0$ is dense in $\mathcal{F}$ under the strict topology, which proves (ii). Point (iii) in turn follows from Theorem 7 and Theorem 9(iii). ∎

4.3 Features for Parameterized Paths

While it can be useful to ignore time-parameterization, there are also many situations where time-parameterization matters, e.g. financial data. The simple observation is that we can record the parameterization of a path by adding time as a state-space coordinate.

---

7. The quotient map $\mathcal{C}^1 \to \mathcal{P}^1$, when $\mathcal{C}^1$ is equipped with the pseudo-metric ρ, is continuous due to the continuity of $\mathrm{S} : \mathcal{C}^1 \to \mathbf{T}_1(V)$ (Lyons and Qian, 2002, Thm. 3.1.3). In particular, the quotient topology on $\mathcal{P}^1$ is *stronger* than the initial topology we consider.

Let $\overline{V} := V \oplus \mathbb{R}$. Observe that the map

$$\mathcal{C}^1(V) \ni x \mapsto \overline{x} \in \mathcal{C}^1(\overline{V}) \text{ with } \overline{x} : [0, T_x] \to \overline{V}, \quad \overline{x}(t) := (x(t), t)$$

realises $\mathcal{C}^1(V)$ as a subset of $\mathcal{C}^1(\overline{V})$. Furthermore, for $x, y \in \mathcal{C}^1(V)$, it holds by strict monotonicity of the second component that $\overline{x} \sim_t \overline{y}$ if and only if $x = y$. Hence, denoting by $\overline{\rho}(x, y)$ the quantity (11) with x, y replaced by $\overline{x}, \overline{y}$, it holds that $(\mathcal{C}^1(V), \overline{\rho})$ is a genuine metric space (which is sensitive to the parameterization of paths).

Corollary 22 *The statement of Theorem 21 holds with $\mathcal{P}^1$ replaced by $(\mathcal{C}^1, \overline{\rho})$ and Φ replaced by the map $\overline{\Phi} : \mathcal{C}^1(V) \to \mathbf{T}_1(\overline{V}), \, x \mapsto \Lambda \circ \mathrm{S}(\overline{x})$.*

Proof The above remarks imply that the map $x \mapsto \overline{x}$ injects $(\mathcal{C}^1(V), \overline{\rho})$ continuously into $\mathcal{P}^1(\overline{V})$, and the conclusion follows from Theorem 21. ∎

A particularly important corollary is that that the robust signature moments characterize laws of stochastic processes.

Corollary 23 *Let μ, ν be probability measures on $\mathcal{C}^1(V)$. Then*

$$\mu = \nu \quad \text{iff} \quad \mathbb{E}_{X \sim \mu}\big[\overline{\Phi}(X)^m\big] = \mathbb{E}_{Y \sim \nu}\big[\overline{\Phi}(Y)^m\big] \quad \forall m \geq 1.$$

Corollary 23 is in stark contrast to classical signature moments, i.e. it does not hold when $\overline{\Phi}(X), \overline{\Phi}(Y)$ are replaced by $\mathrm{S}(\overline{X}), \mathrm{S}(\overline{Y})$; see Appendix B.5.

4.4 Lifting State Space Features to Pathspace Features

So far we have focused on paths that evolve in a linear space. In this section we briefly discuss how one can generalize the robust signature maps from Sections 4.2 and 4.3 into a feature map for paths that evolve in general topological spaces.

Example 7 *Let $\mathcal{X}$ be the space of finite graphs[8] with vertex set $\{1, \ldots, N\}$. Consider the feature map $\varphi : \mathcal{X} \to E := \mathbb{R}^{N \times N}$ that maps a graph to its adjacency matrix. Then $x \to (\Phi \circ \varphi)(x)$ is a natural candidate for a feature map for paths evolving in the space of graphs (an evolving network). Other choices for φ are of course possible and often more sensible, e.g. those arising from kernels and we return to this in Section 6.1.*

Even for paths that evolve in a linear space $\mathcal{X} = \mathbb{R}^d$, precomposition of Φ with a state-space non-linearity φ can be beneficial in terms of efficiency since the state space non-linearity can make it easier to linearize non-linear functions.

Definition 24 *Let $\mathcal{X}$ be a topological space and $\varphi : \mathcal{X} \to V$. Denote with $\mathcal{C}^1_{x_0}(\mathcal{X})$ the subset of paths x in $\cup_{T>0} C([0, T], \mathcal{X})$ such that $x(0) = x_0$ and $\varphi(x) = (\varphi(x(t)))_{t \in [0, T_x]} \in \mathcal{C}^1(V)$. Denote with $\mathcal{P}^1_{x_0}(\mathcal{X}) = \mathcal{C}^1_{x_0}(\mathcal{X})/\sim_t$ the quotient space[9] under tree-like equivalence.*

---

8. Say, directed and weighted graphs without multiple edges; however, any definition of a graph for which an adjacency matrix is well-defined is allowed.
9. Equipped with the initial topology induced by the signature map $\mathrm{S} \circ \varphi : \mathcal{P}^1_{x_0}(\mathcal{X}) \to \mathbf{T}_1(V)$.

Corollary 25 *Suppose $\varphi : \mathcal{X} \to V$ is injective and Λ is a tensor normalization. The statement of Theorem 21 holds with $\mathcal{P}^1(V)$ replaced by $\mathcal{P}^1_{x_0}(V)$ and Φ replaced by the map*

$$\Phi^\varphi : \mathcal{C}^1(\mathcal{X}) \to \mathbf{T}_1(V), \quad \Phi^\varphi := \Phi \circ \varphi .$$

Furthermore, the same statement holds when $\mathcal{P}^1_{x_0}(\mathcal{X})$ is replaced by $(\mathcal{C}^1_{x_0}(\mathcal{X}), \overline{\rho})$ and Φ^φ is replaced by $\overline{\Phi}^\varphi := \Lambda \circ S(\overline{\varphi(x)})$.

Proof Injectivity of φ implies that for all $x, y \in \mathcal{C}^1_{x_0}(\mathcal{X})$, $\varphi(x) \sim_t \varphi(y)$ if and only if $x \sim_t y$. All the desired claims now follow from Theorem 21 and Corollary 22. ∎

Applied with $\mathcal{X} = V$ and $\varphi = \mathrm{id} : \mathcal{X} \to V$, this just recovers the feature map $\Phi^{\mathrm{id}} = \Phi$ for $\mathcal{P}^1$. Lifts of $\mathcal{X}$-valued sequences were introduced for sequence kernels in (Kiraly and Oberhauser, 2019).

5. Robust Features for Rough Paths

In this section we enlarge the domain of the feature map Φ (resp. $\overline{\Phi}$) to paths of unbounded 1-variation. This allows to include important classes of stochastic processes such as stochastic differential equations, semimartingales, Markov processes, and Gaussian processes. The main obstacle is that the iterated integrals $\int dx^{\otimes m}$ can no longer be defined by Riemann–Stieltjes integration since the trajectories of x are too irregular. Rough paths theory provides a complete integration theory for a large class of processes; for the special case of (semi)martingales, the theory agrees with Itô integration. We focus here on the main ideas but give a detailed account and examples in Appendix B.

Rough path theory provides a family of spaces $(\mathcal{C}^p)_{p \geq 1}$ called geometric p-rough paths over V. These spaces satisfy the inclusions $\mathcal{C}^p \subset \mathcal{C}^q$ for $p \leq q$, and larger values of p allow for paths with "rougher" trajectories. For every $p \geq 1$, there exists a map $\mathcal{C}^p \to \mathbf{T}_1(V)$ which has the same properties as the signature map S from Section 4. In fact, for $p = 1$, $\mathcal{C}^1$ is just the space of absolutely continuous paths and we recover the setting of Section 4. Similar to before, we define the space $\mathcal{P}^p = \mathcal{C}^p / \sim_t$, where $\sim_t$ denotes tree-like equivalence, which we equip with the topology induced by embedding $\mathcal{P}^p \to \mathbf{T}_1(V)$. The following is a significant generalisation of Theorem 21 to large classes of processes, see Example 8. The proof follows however, analogous to the proof of Theorem 21.

Theorem 26 *Let $p \geq 1$ and $\Lambda : \mathbf{T}_1(V) \to \mathbf{T}_1(V)$ a tensor normalization. The map*

$$\Phi : \mathcal{P}^p \to \mathbf{T}_1(V) , \quad \Phi = \Lambda \circ S$$

(i) is a continuous injection from $\mathcal{P}^p$ into a bounded subset of $\mathbf{T}_1(V)$,

(ii) is universal to $\mathcal{F} := C_b(\mathcal{P}^p, \mathbb{R})$ equipped with the strict topology,

(iii) is characteristic to the space of finite regular Borel measures on $\mathcal{P}^p$.

Moreover, $\forall \ell \in \mathbf{T}(V)'$ the map $\mu \to \langle \ell, \mathbb{E}_{X \sim \mu}[\Phi(X)] \rangle$ is B-robust. The same statement holds when $\mathcal{P}^p$ is replaced by $\mathcal{C}^p$ and Φ is replaced by $\overline{\Phi}(\mathbf{x})$, see Appendix B.3.

18

6. A Computable Metric for Laws of Stochastic Processes

Sections 4 and 5 provide us with a robust feature map Φ (resp. $\overline{\Phi}$) such that its expected value $\mathbb{E}_{X \sim \mu}[\Phi(X)]$ characterizes the law μ of the stochastic process X [with the option to ignore parameterization]. In this section we use this to derive a metric on the space of laws of stochastic processes and study the topology it induces.

A natural distance for probability measures on any topological space $\mathcal{X}$ is given by fixing a class of functions $\mathcal{G} \subset \mathbb{R}^{\mathcal{X}}$ and define the *maximum mean discrepancy* (MMD) between μ and ν as

$$d(\mu, \nu) := \sup_{f \in \mathcal{G}} |\mathbb{E}_{X \sim \mu}[f(X)] - \mathbb{E}_{Y \sim \nu}[f(Y)]| \ .$$

If $\mathcal{G}$ is sufficiently large, then d defines a metric between probability measures, see (Rachev, 1991; Müller, 1997). However, the supremum makes it hard to compute or estimate $d(\mu, \nu)$. An insight from kernel learning is that if $\mathcal{G}$ is the unit ball of a RKHS $(\mathcal{H}, k)$, then the reproducing property implies

$$d^2(\mu, \nu) = \mathbb{E}[k(X, X')] - 2\mathbb{E}[k(X, Y)] + \mathbb{E}[k(Y, Y')] \ , \tag{15}$$

where $X, X' \sim \mu$ and $Y, Y' \sim \nu$ are pairwise independent. The above can be simply estimated from finite samples of μ and ν provided k is cheap to evaluate.

We are interested in the case when $\mathcal{X} = \mathcal{C}^p$ is the space of p-rough paths and μ and ν are the laws of stochastic processes $X = (X_t)$ and $Y = (Y_t)$. The robust signature feature map Φ gives us a natural candidate to define a kernel $k(x, y) := \langle \Phi(x), \Phi(y) \rangle$. A well-known disadvantage of MMDs is that they metrize a topology that is usually not comparable (i.e. neither weaker nor stronger) to weak convergence, see (Rachev, 1991; Müller, 1997). Progress on this question was made recently in (Simon-Gabriel and Schölkopf, 2018; Simon-Gabriel et al., 2020) in the case of locally compact spaces $\mathcal{X}$ building on (Sriperumbudur, 2016; Sriperumbudur et al., 2011). However, already the space of absolutely continuous paths $\mathcal{C}^1$ is not locally compact.

6.1 Kernel Learning

We briefly recall notation and terminology from kernel learning (Cucker and Smale, 2002; Schölkopf and Smola, 2002; Berlinet and Thomas-Agnan, 2011). Throughout this section, let $\mathcal{X}$ be a topological space. Let us fix a feature map $\mathcal{X} \to \mathbb{R}^{\mathcal{X}}$, $x \mapsto k_x$, and denote by $\mathcal{H}_0 \subset \mathbb{R}^{\mathcal{X}}$ the linear span of $\{k_x : x \in \mathcal{X}\}$. Under the condition that $k(x, y) := k_x(y)$ is a symmetric positive definite kernel (which we assume henceforth), the completion of $\mathcal{H}_0$ under the inner product $\langle k_x, k_y \rangle := k(x, y)$ is the RKHS of k. We further fix a locally convex TVS $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ containing $\mathcal{H}_0$ for which the inclusion map $\mathcal{H}_0 \hookrightarrow \mathcal{F}$ is continuous.

Definition 27 (Simon-Gabriel and Schölkopf (2018)) *We say that the kernel* k *is*

- *universal to* $\mathcal{F}$ *if the kernel embedding* $\iota = \mathrm{id} : \mathcal{H}_0 \hookrightarrow \mathcal{F}$, $f \mapsto f$, *is dense,*

- *characteristic to* $\mathcal{F}'$ *if the kernel mean embedding* $\mu : \mathcal{F}' \to \mathcal{H}_0'$, $D \mapsto D|_{\mathcal{H}_0}$ *is injective.*

Proposition 28 (Simon-Gabriel and Schölkopf (2018)) *The kernel embedding* ι *and the kernel mean embedding* μ *are transpose:* $\iota^\star = \mu$ *and* $\mu^\star = \iota$. *Moreover,* k *is universal to* $\mathcal{F}$ *iff* k *is characteristic to* $\mathcal{F}'$.

A common way to construct the feature map $x \mapsto \mathrm{k}_x$ is through another map $\Phi : \mathcal{X} \to E$ taking values in an inner product space $(E, \langle \cdot, \cdot \rangle)$. Then one defines $\mathrm{k}_x(y) := \mathrm{k}(x, y) := \langle \Phi(x), \Phi(y) \rangle$. We recall the important details of this construction in Appendix C. As expected, universality and characteristicness of Φ and k are equivalent.

Proposition 29 *Consider a map $\Phi : \mathcal{X} \to E$ into an inner product space $(E, \langle \cdot, \cdot \rangle)$ such that $\mathrm{k}(x, y) = \langle \Phi(x), \Phi(y) \rangle$. Then*

- *Φ is universal to $\mathcal{F}$ iff k is universal to $\mathcal{F}$,*

- *Φ is characteristic to $\mathcal{F}'$ iff k is characteristic to $\mathcal{F}'$.*

Proof This follows from Proposition 43 in Appendix C. ∎

6.2 The Robust Signature Kernel and its MMD

The feature map Φ defined in Section 5 is universal and characteristic for (unparameterized) paths that evolve in a Banach space V. It takes values in the feature space $\mathbf{T}(V)$ that is itself a Banach space. In particular, Φ can be applied to (unparameterized) paths that evolve in a Hilbert space $(H, \langle \cdot, \cdot \rangle)$. In this case the feature space $\mathbf{T}(H)$ is also a Hilbert space, see Section 3.1. This allows to take an inner product in the feature space which immediately gives rise to a kernel on $\mathcal{P}^p(H)$ that is also universal and characteristic.

Theorem 30 *Let $p \geq 1$ and $\Lambda : \mathbf{T}_1(H) \to \mathbf{T}_1(H)$ a tensor normalization. Define*

$$\mathrm{k} : \mathcal{P}^p(H) \times \mathcal{P}^p(H) \to \mathbb{R}, \quad (x, y) \mapsto \langle \Phi(x), \Phi(y) \rangle$$

where $\Phi = \Lambda \circ \mathrm{S}$ is the normalized signature map. We call k the robust signature kernel and denote with

$$d_{\mathrm{k}}(\mu, \nu) := \sup_{f \in \mathbf{T}(H) : \|f\| \leq 1} \left| \int f(x)\mu(dx) - \int f(y)\nu(dy) \right|$$

the associated MMD. Then k is a bounded, continuous, positive definite kernel and

  (i) k is universal to $C_b(\mathcal{P}^p(H), \mathbb{R})$ equipped with the strict topology,

  (ii) k is characteristic to finite, signed Borel measures on $\mathcal{P}^p(H)$,

  (iii) d_{k} is a metric on the space of finite, signed Borel measures on $\mathcal{P}^p(H)$, and

$$d_{\mathrm{k}}^2(\mu, \nu) = \mathbb{E}[\mathrm{k}(X, X')] - 2\mathbb{E}[\mathrm{k}(X, Y)] + \mathbb{E}[\mathrm{k}(Y, Y')]$$

  where $X, X' \sim \mu$ and $Y, Y' \sim \nu$ are pairwise independent,

  (iv) for probability measures on $\mathcal{P}^p(H)$, convergence in d_{k} does not imply weak convergence. Suppose now that H is finite dimensional. Then weak convergence implies convergence in d_{k}. Furthermore, if $\mathcal{M}$ is a set of probability measures on $\mathcal{P}^p(H)$ which is compact under weak convergence, then d_{k} induces the topology of weak convergence on $\mathcal{M}$.

The same statement holds when $\mathcal{P}^p(H)$ is replaced by $(\mathcal{C}^p(H), \overline{\rho}_p)$ and Φ in $\mathrm{k}(x, y) = \langle \Phi(x), \Phi(x) \rangle$ is replaced by $\overline{\Phi}(\mathbf{x}) := \Lambda \circ \mathrm{S}(\overline{\mathbf{x}})$.

Proof All claims, except (iv), follow from combining Theorem 26 and Proposition 29. The proof of (iv) is more involved and follows from Proposition 40 in Appendix B.4. ∎

6.3 Lifting State Space Kernels to Pathspace Kernels

In Section 4.4 we discussed how precomposing the feature map Φ with a state space non-linearity φ can be beneficial. The same applies to the robust signature kernel k.

Proposition 31 *Let (H, κ) be a RKHS on $\mathcal{X}$ such that $\varphi : \mathcal{X} \to H$, $\varphi(x) := \kappa(x, \cdot)$, is injective[10] and continuous. For $x \in C([0, T], \mathcal{X})$ denote by $x \mapsto \kappa_x := (\varphi(x(t)))_t \in C([0, T], H)$ the lift of x to a path evolving in H. We call[11]*

$$\mathrm{k}^\kappa : \mathcal{P}^1(\mathcal{X}) \times \mathcal{P}^1(\mathcal{X}) \to \mathbb{R}, \quad (x, y) \mapsto \mathrm{k}(\kappa_x, \kappa_y)$$

the robust signature lift of the kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Then k^κ is a bounded, continuous, positive definite kernel and, using the notation from Definition 24, for any $x_0 \in \mathcal{X}$

(i) k^κ *is universal to $C_b(\mathcal{P}_{x_0}^1(\mathcal{X}), \mathbb{R})$ equipped with the strict topology,*

(ii) k^κ *is characteristic to finite, signed Borel measures on $\mathcal{P}_{x_0}^1(\mathcal{X})$,*

(iii) d_{k^κ} *is a metric on the space of finite, signed Borel measures on $\mathcal{P}_{x_0}^1(\mathcal{X})$.*

The same statement holds when $\mathcal{P}_{x_0}^1(\mathcal{X})$ is replaced by $(\mathcal{C}_{x_0}^1(\mathcal{X}), \overline{\rho})$ and Φ (implicit in the definition of k) is replaced by $\overline{\Phi}(x) := \Lambda \circ \mathrm{S}(\overline{x})$.

If $\mathcal{X}$ is itself a Hilbert space, the choice $\kappa(\cdot, \cdot) = \langle \cdot, \cdot \rangle$ corresponds to $\varphi = \mathrm{id}$ and recovers the signature kernel k from Theorem 30, $\mathrm{k}^{\langle \cdot, \cdot \rangle} = \mathrm{k}$.

6.4 Computing the Signature Kernel

We now turn to the computational aspects of the robust signature kernel k^κ and associated MMD. In practice, we are not given a path sample $x = (x(t))_{t \in [0,T]}$, but a sequence sample $x^\pi = (x(t_i))_i$ that measures the values of x at times t_i in a partition $\pi = \{0 \le t_1 < \cdots \le T\}$ of $[0, T]$. This partition π might even change from path to path in a given dataset. Hence, instead of working with probability measure on paths we work with probability measures on sequences. Below we define the discrete robust signature (resp. kernel resp. MMD) on the domain of sequences and then quantify how far these are from the robust signature Φ^φ (resp. kernel k^κ resp. MMD d_{k^κ}) that is defined on the domain of paths. Further, we show that the algorithms of Kiraly and Oberhauser (2019) can be used to compute these quantities efficiently.

For the rest of this section, we fix an integer $M \ge 1$, a RKHS (H, κ) on $\mathcal{X}$, and a K-Lipschitz function ψ as in Corollary 15 with associated tensor normalization Λ.

---

10. φ is injective whenever κ is characteristic to, say, the space of finite, signed Borel measures on $\mathcal{X}$.
11. Note that k in $\mathrm{k}(\kappa_x, \kappa_y)$ is the signature kernel on $\mathcal{P}^1(H)$ from Theorem 30.

Definition 32 *Denote by $\mathcal{X}^+ = \bigcup_{\ell \geq 0} \mathcal{X}^\ell$ the space of sequences in $\mathcal{X}$ of arbitrary length. Define*

$$\mathrm{k}_M^{\kappa,+} : \mathcal{X}^+ \times \mathcal{X}^+ \to \mathbb{R} \,, \quad (x,y) \mapsto \langle \Phi_M^{\mathrm{linear}}(\kappa_x), \Phi_M^{\mathrm{linear}}(\kappa_y) \rangle \tag{16}$$

where, for $x = (x_0, \ldots, x_\ell)$, κ_x is the H-valued sequence $(\kappa(x_0, \cdot), \ldots, \kappa(x_\ell, \cdot))$ and[12]

$$\Phi_M^{\mathrm{linear}} : \mathcal{X}^+ \to \prod_{m=0}^{M} H^{\otimes m}, \quad \Phi_M^{\mathrm{linear}} = \Lambda \Big(\prod_{i=1}^{\ell} (1 + \kappa(x_i, \cdot) - \kappa(x_{i-1}, \cdot)) \Big) \,. \tag{17}$$

As we show in Proposition 44 of Appendix C.1, $\mathrm{k}_M^{\kappa,+} : \mathcal{X}^+ \times \mathcal{X}^+ \to \mathbb{R}$ is a bounded, positive semidefinite kernel for $\mathcal{X}^+$. The reason why $\mathrm{k}_M^{\kappa,+}$ is of practical interest is that it approximates k^κ with an explicit convergence rate. More precisely. for $x, y \in \mathcal{C}^1(\mathcal{X})$ and partitions $\pi \subset [0, T_x], \pi' \subset [0, T_y]$,

$$\left| \mathrm{k}^\kappa(x, y) - \mathrm{k}_M^{\kappa,+}(x^\pi, y^{\pi'}) \right| \leq C \big(\max_{t_i \in \pi} \|\kappa_x\|_{1-\mathrm{var};[t_{i-1}, t_i]}^{1/2} + \max_{t_i \in \pi'} \|\kappa_y\|_{1-\mathrm{var};[t_{i-1}, t_i]}^{1/2} \big) \,,$$

where x^π (resp. $y^{\pi'}$) denote the sequences given by sampling x, y along π (resp. π'), and where C depends only on $\|\psi\|_\infty, K, \|\kappa_x\|_{1-\mathrm{var}}$, and $\|\kappa_y\|_{1-\mathrm{var}}$. A similar approximation result holds for the associated MMDs d_{k^κ} and $d_{\mathrm{k}_M^{\kappa,+}}$. See again Proposition 44 for a precise statement.

The following result provides a bound on the computation cost of $\mathrm{k}_M^{\kappa,+}$.

Proposition 33 *The kernel $\mathrm{k}_M^{\kappa,+}(x^\pi, y^{\pi'})$ can be evaluated in $O((|\pi|^2 + |\pi'|^2)(c + M) + q)$ time and $O(|\pi|^2 + |\pi'|^2 + M + r)$ memory, where $|\pi|$ is the number of time points in π, c is the cost of one evaluation of the kernel κ, and q and r are the total time and memory costs respectively of a single evaluation of ψ and of finding the unique non-negative root of a polynomial $P(\lambda) = \sum_{m=0}^{M} a_m \lambda^{2m}$ with $a_0 \leq 0 \leq a_1, \ldots, a_M$.*

Proof Denote $\mathrm{S}_M^{\mathrm{linear}}(h) := \prod_{i=1}^{\ell} (1 + h_i - h_{i-1})$. The sequence of norms $\|(\mathrm{S}_M^{\mathrm{linear}}(\kappa_{x^\pi})^m\|_{H^{\otimes m}}$, $m = 0, \ldots, M$, can be computed in $O(|\pi|^2(c + M))$ time and $O(|\pi|^2 + M)$ memory due to (Kiraly and Oberhauser, 2019, Alg. 3). In the same way, the sequence of inner products $(\langle \mathrm{S}_M^{\mathrm{linear}}(\kappa_{x^\pi})^m, \mathrm{S}_M^{\mathrm{linear}}(\kappa_{y^{\pi'}})^m \rangle_{H^{\otimes m}})_{m=0}^{M}$ can be computed in $O(|\pi||\pi'|(c + M))$ time and $O(|\pi||\pi'| + M)$ memory. We now claim that, for $\mathbf{s}, \mathbf{t} \in \prod_{m=0}^{M} H^{\otimes m}$, once the three sequences $(\|\mathbf{s}^m\|)_{m=0}^{M}, (\|\mathbf{t}^m\|)_{m=0}^{M}$, and $(\langle \mathbf{s}^m, \mathbf{t}^m \rangle)_{m=0}^{M}$ are given, we can compute the normalized inner product $\langle \Lambda(\mathbf{s}), \Lambda(\mathbf{t}) \rangle$ in $O(q)$ time and $O(r)$ memory, which will finish the proof. Indeed, let $\lambda(\mathbf{t}) \geq 0$ be as in Proposition 14. Then $\langle \Lambda(\mathbf{s}), \Lambda(\mathbf{t}) \rangle = \sum_{m=0}^{M} \lambda(\mathbf{s})^m \lambda(\mathbf{t})^m \langle \mathbf{s}^m, \mathbf{t}^m \rangle$, so it suffices to compute $\lambda(\mathbf{s})$ and $\lambda(\mathbf{t})$. These can be computed in the claimed time and memory since $\lambda(\mathbf{t})$ is the unique non-negative root of the polynomial $P(\lambda) := \|\delta_\lambda \mathbf{t}\|^2 - \psi(\|\mathbf{t}\|) = \left(\sum_{m=0}^{M} \lambda^{2m} \|\mathbf{t}^m\|^2 \right) - \psi(\|\mathbf{t}\|)$. ∎

The quadratic complexity in the sequence length $|\pi|$ can be reduced to linear complexity by using the simultaneous dynamic programming and low-rank algorithm from (Kiraly and Oberhauser, 2019, Algorithm 5, Section 6); however, although the low-rank assumption works well in practice, it is unclear how to justify it theoretically in the context of paths and signatures.

---

12. We use the shorthand notation $1 + \kappa(x_i, \cdot) - \kappa(x_{i-1}, \cdot) := (1, \kappa(x_i, \cdot) - \kappa(x_{i-1}, \cdot), 0, \ldots, 0) \in \prod_{m=0}^{M} H^{\otimes m}$ and $\prod_{i=1}^{\ell} x_i$ denotes the product $x_1 \otimes x_2 \otimes \cdots \otimes x_\ell$ in the tensor algebra, see Appendix A.

7. Application: Two-sample Tests for Stochastic Processes

We apply the MMD metric developed in the previous sections to the two-sample testing problem

$$H_0 : \mu = \nu \text{ against the alternative } H_1 : \mu \neq \nu \tag{18}$$

where $\mu := \mathrm{Law}(X) := \mathbb{P} \circ X^{-1}$ and $\nu := \mathrm{Law}(Y) := \mathbb{P} \circ Y^{-1}$ denote the laws of the stochastic processes X and Y. We compare the resulting MMD test against classical multi-variate tests as well as MMD tests using classical kernels for vector-valued data.

7.1 Test statistics

Denote with $X = (X(t_1), \dots, X(t_L)), Y = (Y(s_1), \dots, Y(s_L))$ two independent, discrete-time stochastic processes that evolve in $\mathbb{R}^d$. We are given m (resp. n) i.i.d. samples from X (resp. Y), which we denote with $\mathbf{X} = \{X_1, \dots, X_m\}$ and $\mathbf{Y} = \{Y_1, \dots, Y_n\}$. That is, each X_i is a sequence of L vectors in $\mathbb{R}^d$. The probability of falsely rejecting the null is called the *Type I error*, and similarly the probability of falsely accepting the null is called the *Type II error*. If the Type I error can be bounded from above, uniformly over all $\mathbb{P}$ under which X and Y are independent, by a constant α, then we say that the test is of *significance level* α. Given a statistic $T(\mathbf{X}, \mathbf{Y})$ one can construct a simple permutation test: under the null H_0 one can sample from the distribution of $T(\mathbf{X}, \mathbf{Y})$ by sampling uniformly at random from the $(m+n)!$ permutations π of $\{1, \dots, m+n\}$ and evaluate

$$T(Z_{\pi(1)}, \dots, Z_{\pi(m)}, Z_{\pi(m+1)}, \dots, Z_{\pi(m+n)})$$

where $\mathbf{Z} = (X_1, \dots, X_m, Y_1, \dots, Y_n)$. In particular, this allows to estimate the 95% quantile of $T(\mathbf{X}, \mathbf{Y})$ under the null H_0. Hence, if $T(\mathbf{X}, \mathbf{Y})$ is not in this 95% quantile, we can reject the null with significance level 5%. For background on testing and (permutation) test statistics we refer to (Lehmann and Romano, 2006, Chapter 15).

MMD statistics. Using the sequence signature kernel $\mathrm{k}_M^{\kappa,+}$ from Definition 32, we can use the MMD

$$d_{\mathrm{k}_M^{\kappa,+}}(\mu, \nu)^2 = \mathbb{E}[\mathrm{k}_M^{\kappa,+}(X, X')] - 2\mathbb{E}[\mathrm{k}_M^{\kappa,+}(X, Y)] + \mathbb{E}[\mathrm{k}_M^{\kappa,+}(Y, Y')] \tag{19}$$

where X, X' (resp. Y, Y') denote pairwise independent sequences sampled from μ (resp. ν). The unbiased (quadratic complexity) estimator from (Gretton et al., 2009, 2012) turns (19) into a test statistic $T(\mathbf{X}, \mathbf{Y})$ and combining this with the results of Section 6 allows to efficiently estimate (19). Moreover, we can do the same by replacing the robust sequence signature kernel $\mathrm{k}_M^{\kappa,+}$ in (19) with any kernel κ for vector-valued data by flattening sequence samples; e.g. if X evolves in the state space $\mathbb{R}^d$, we identify the sequence of vectors $(X(t_1), \dots, X(t_L))$ as a single vector in $\mathbb{R}^{d \cdot L}$. The choices for the kernel κ we used (both in the signature MMD and for the vector kernel MMD of the flattened sequence) were the linear kernel $\kappa(x, y) = \langle x, y \rangle$, the RBF kernel $\kappa(x, y) = \exp(-\|x - y\|^2/(2\sigma^2))$, and the Laplace kernel $\kappa(x, y) = \exp(-\|x - y\|/\sigma)$ for a parameter $\sigma > 0$. Kernels depend on parameters: for the signature MMD these parameters are the truncation level M, the parameters a, C that determine the tensor normalization (Example 4), and the parameters of the kernel κ; for the classical kernel choices of κ the only parameter is the length scale σ.

Classic multi-variate statistics. By flattening a sequence observations into a long vector (as described above), one can also use classical multi-variate test statistics $T(\mathbf{X}, \mathbf{Y})$. Many of these classical tests like Biau-Györfi scale poorly and the high-dimensionality of our testing problem makes it impossible to use them; see (Gretton et al., 2012) for more on this. As in (Gretton et al., 2012) we settled for the Friedman–Rafsky multivariate Wald–Wolfowitz test statistic from (Friedman and Rafsky, 1979), (Wolf), and the Hotelling t-test statistic, (Hotelling), since they are still computable in moderately high-dimension. However, even here, computing these tests requires building a graph for (Wolf) and matrix inversion for (Hotelling). This makes them scale at least cubically in dL and thus are only applicable if both, the dimension of the state space d and the sequence length L, are small.

7.2 Data

As data sources we use the multi-variate time series datasets from UAE & UCR, see (Bagnall et al., 2018) for a detailed description. This archive consists of 30 time series datasets from various domains, including speech, motion, handwriting, robotics, etc. The state space $\mathbb{R}^d$ of these datasets varies between low-dimensional $d = 2$ (e.g. Pendigits) to high-dimensional $d = 1345$ (e.g. DuckDuckGeese).

For each time-series dataset we define μ (resp. ν) as follows: under the null H_0 a sample from μ (resp. ν) is generated by choosing uniformly at random a time-series $W = (W(t_1), \ldots, W(t_{L_0}))$ from this dataset and double the sequence length by concatenation; that is $X = (W(t_1), \ldots, W(t_{L_0}), W(t_1), \ldots, W(t_{L_0}))$. Under the alternative H_1, a sample from μ is generated as above, but a sample from ν is generated by concatenating in reverse, $Y = (W(t_1), \ldots, W(t_{L_0}), W(t_{L_0}), \ldots, W(t_1))$. Despite its simplicity, we will see that this problem is challenging for many statistics, since a statistic needs to capture the order structure to achieve a high test power. In our experiments, we varied the length $L_0 \in \{10, 100, 200\}$ by truncating after the first L_0 entries of the time-series in the original dataset. Further, we restricted to the case of balanced samples, $m = n$, and varied the number of samples $m \in \{30, 70, 200\}$. This allows to also compare the sensitivity of the different statistics to length and sample size.

7.3 Implementation Details

As base for our signature MMD computations we use Csaba Toth's KSig library[13] that provides efficient GPU implementations of signature kernels in Python. As base for general kernel MMD two-sample permutation test, we built on code from the DS3 summer school course "Representing and comparing probabilities with kernels" provided by H. Strathmann and D. Sutherland[14] but we rewrote much of it to make it compatible with the KSig package. To find the non-negative root of the polynomial $P(\lambda)$ from Proposition 33 we used an optimization of Brent's method (Brent, 1973, Ch. 3-4) implemented as `optimize.brentq` in the `SciPy` package (Jones et al., 2001–).

---

13. `https://github.com/tgcsaba/KSig`
14. `https://github.com/karlnapf/ds3_kernel_testing`

7.4 Discussion of results

For each test statistic we used the permutation test with significance level 5% and repeated the following for each dataset 20 times: (i) Generate samples **X**,**Y** under the null H_0 and record the number of false rejections of the null H_0 (*Type I error*). (ii) Generate samples **X**,**Y** under the alternative H_1 and record the number of false acceptances of the null H_0 (*Type II error*). The power of the test—the probability of rejecting the null H_0 when H_1 is true—then determines which test performs best; the closer to 1 the better. As mentioned above for classical statistics the above was computationally only possible when the effective dimension is small hence these are only included in the tables when feasible.

Table 2 and Table 3 show the empirical Type I error and 1 minus the Type II error (i.e. the power of the test) for the different statistics for sequence length $L_0 = 10$ and $m = 30$. Table entries are rounded to two decimal digits and the most powerful test for each dataset is highlighted in bold. The remaining results for the Type II error with varying sequence length L_0 and sample size m are shown in Tables 4, 5, and 6. For brevity we do not include tables with the Type I error for these other regimes since, by construction of the permutation test, these errors converge to a number less than 5% as m increases (that is, they all look very similar to Table 2, just more entries are less or equal to 0.05).

The experiments show that, as one expects, the linear (inner product) kernel MMD does poorly overall and generally fails to distinguish two different distributions. The standard MMDs with RBF and Laplace kernel outperform the linear kernel, but in turn rarely reach the test power of any of the signature MMDs. As the length increases and thus samples carry more distinguishable structure, the performance of all MMDs improves. However, qualitatively the same behaviour remains. The same is true when the number of samples m increases; see Tables 3-6. Among the classical statistics, the Hotelling test statistic is not competitive, even for small sequence length. The Friedman-Rafsky multivariate version of the Wald-Wolfowitz test statistic performs well overall and is only beaten by the linear Signature MMD. However, this test comes at a huge computational cost even for short sequences since it involves computing a minimal spanning tree. Similarly Hotelling involves a matrix inversion. This is the reason why can report these two, Hotelling and Wolf, only for the shortest sequence length $L_0 = 10$ and even there only for time-series that evolve in relatively low-dimensional state spaces (these experiments took several days on a multicore machine,[15] as opposed for the MMD experiments which took about an hour).

Overall, these results show that signature MMDs efficiently exploit the sequence structure whereas classical MMDs need much more data to reach a similar test power. A more surprising observation is that for the signature MMDs, the linear kernel does better than non-linear kernels. This is in stark contrast to supervised time-series classification where non-linear signature kernels strongly outperform the linear signature kernel as in (Toth and Oberhauser, 2020). We believe the reason is that the additional parameter selection for the non-linear kernel introduces additional variance, but unlike for the non-signature MMDs, the linear signature MMD is already expressive enough so that the additional non-linearities only result in more variance. This trade-off might change considerably if one switches to other data resp. hypothesis tests.

---

15. All experiments were run on a machine with 36 cores (2 x 18 Core with HT), 2.6/3.9 GHz, 1536GB RAM.

Related to this is also the question of kernel parameter choice. All our tables show the results when the approach in (Fukumizu et al., 2009) is applied for the kernel parameter choice. We also experimented with the median heuristic as well as (Sutherland et al., 2016) that uses a part of the data for training to directly minimize the Type II error. The median heuristic was not competitive and the price for learning the parameters is that the training data can not be used for the test itself and the choice of the train-test split is itself a heuristic. In our experiments this choice had big effects from dataset to dataset so that ultimately (Fukumizu et al., 2009) yielded overall the best and consistent performance for all MMDs. Nevertheless, improving the parameter choice for MMD tests is an ongoing research effort and any progress can potentially improve the results of all MMDs. The parameter selection resulted in a value for M that was less or equal to 5 for all signature MMDs. Furthermore, for the signature MMDs build on top of a non-linear kernel κ, the value of M was typically even lower, often equal to 2. This is in line with the discussion in Section 4.4 and Section 6.3 that state-space non-linearities help to linearize so that a few iterated integrals suffice. In contrast, for the signature MMD build on top of the linear kernel, large values of M could be in principle required. However, the price of using iterated integrals $\int dX^{\otimes M}$ for large values of M is that their variance increases since we only have access to an empirical measure that approximates their expectation. This explains why for the linear signature MMD, the optimal parameter M is typically higher than for signature MMDs with non-linear κ but still relatively low. The other parameters of the signature MMD a, C determine the tensor normalization, Example 4. For simplicity we kept both fixed ($a = 1, C = 10^3$) since the theoretical guarantees apply to any choice and the signature MMD already outperform the baselines. However, learning them could potentially improve the signature MMD results further. The parameter for the length scale σ varied over several orders of magnitude from dataset to dataset, for both signature MMDs and the other MMDs.

Finally, we note that one of the biggest strengths of the signature MMD is that it applies to sequences of different length. None of the other statistics that identify a sequence as a long vector can be used here since the vectors that result from flattening the sequence would not have the same dimension. To test how much this influences the signature MMDs, we drew for each sequence sample a random integer k between 0 and $0.3 \cdot L_0$ and then deleted uniformly at random k sequence entries. The result is that each datasets contains sequences of varying length. The signature MMDs can be applied without modifications and we repeated the above experiments. The performance slightly deteriorates but remains very close to the signature MMD results given in the tables, hence we omit it for brevity.

To sum up, we recommend to try both, the linear signature MMD and RBF signature MMD, for testing no matter what regime (small/large samples, small/large length, small/large state space).

	MMDs			Signature MMDs			Classic Statistics	
	Linear	RBF	Laplace	Linear	RBF	Laplace	Hotelling	Wolf
ArticularyWordRecognition	0.1	0.05	0.1	0.05	0	0	0.05	0.2
AtrialFibrillation	0.1	0.05	0.05	0	0	0.1	0.1	0.1
BasicMotions	0.05	0	0.1	0.05	0.1	0.1	0.08	0
CharacterTrajectories	0	0.1	0.05	0	0.1	0.05	0.1	0.15
Cricket	0.05	0	0.1	0.1	0.05	0	0.1	0.05
DuckDuckGeese	0.15	0.15	0.05	0.05	0	0.1	-	-
ERing	0	0.05	0.1	0.05	0	0	0.05	0.1
EigenWorms	0.05	0.05	0	0	0.05	0.1	0	0
Epilepsy	0.1	0.05	0.1	0.1	0.05	0	0	0.15
EthanolConcentration	0	0	0.05	0.1	0	0	0.05	0
FaceDetection	0.05	0	0.05	0.05	0	0.1	-	-
FingerMovements	0	0.1	0	0.1	0	0.05	-	-
HandMovementDirection	0	0.05	0	0.05	0	0.05	0	0.05
Handwriting	0.15	0	0.05	0	0.1	0.1	0.1	0.05
Heartbeat	0	0.05	0.15	0	0.1	0.05	-	-
InsectWingbeat	0.05	0.25	0.05	0.1	0	0	-	-
JapaneseVowels	0.25	0	0.05	0	0	0.05	0.05	0.1
LSST	0.05	0.1	0.15	0.05	0.1	0.05	0	0.15
Libras	0.1	0.05	0.1	0	0.1	0.05	0.1	0.1
MotorImagery	0.05	0.1	0.1	0.1	0.05	0.05	-	-
NATOPS	0.05	0.05	0.1	0	0.05	0	-	-
PEMS-SF	0.05	0.05	0	0.05	0.05	0.05	-	-
PenDigits	0.05	0	0.15	0	0	0.1	0.05	0.1
PhonemeSpectra	0.05	0.05	0.15	0.1	0.05	0.05	0.05	0
RacketSports	0	0.05	0.05	0.2	0	0.1	0.1	0
SelfRegulationSCP1	0.15	0.05	0.1	0	0.1	0	0.05	0.05
SelfRegulationSCP2	0.05	0.05	0.05	0	0.05	0.15	0.05	0.1
SpokenArabicDigits	0	0.15	0.05	0.15	0.1	0	0	0.05
StandWalkJump	0.05	0.1	0	0.1	0.15	0.1	0	0.15
UWaveGestureLibrary	0.05	0	0.05	0.05	0	0.05	0.05	0.05

Table 2: Type I error for sequence length $L_0 = 10$ and $m = n = 30$ samples.

	MMDs			Signature MMDs			Classic Statistics	
	Linear	RBF	Laplace	Linear	RBF	Laplace	Hotelling	Wolf
ArticularyWordRecognition	0	0	0.2	**0.55**	0.25	0.3	0.05	0.3
AtrialFibrillation	0.2	0.2	0.15	0.7	0.65	**0.85**	0.15	0.8
BasicMotions	0.2	0.45	0.45	**1**	**1**	**1**	0.4	**1**
CharacterTrajectories	0.3	**1**	**1**	**1**	**1**	**1**	0.2	**1**
Cricket	0.4	0.9	0.95	**1**	**1**	**1**	**1**	**1**
DuckDuckGeese	0.05	0.05	0	**0.9**	0.45	0.2	-	-
ERing	0.05	0.45	0.3	**1**	0.9	0.8	0.3	0.95
EigenWorms	0	0.2	0.35	**1**	0.15	0.15	0.1	0.15
Epilepsy	0.1	0.15	0.1	0.35	0.1	0.1	0.1	**0.7**
EthanolConcentration	0.05	0.05	0.1	**0.3**	0	0.05	0	**0.3**
FaceDetection	0	**0.75**	0.15	0.2	0.05	0.05	-	-
FingerMovements	0.3	0.65	0.5	**1**	0.15	0.05	-	-
HandMovementDirection	0.05	0.05	0.05	**0.95**	0.15	0.1	0.05	0.3
Handwriting	**1**	**1**	**1**	**1**	**1**	**1**	0.8	**1**
Heartbeat	0.05	0	0.1	**0.25**	0	0.05	-	-
InsectWingbeat	0.05	0	0.05	0.1	**0.2**	0.1	-	-
JapaneseVowels	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
LSST	0.1	0.1	0.05	0.35	0.05	0.05	0.14	**1**
Libras	0.05	0.05	0.2	0.15	0.25	0.15	0.2	**0.95**
MotorImagery	0.05	0.05	0.05	**1**	0.7	0.35	-	-
NATOPS	0.1	0.15	0.15	**0.65**	0.15	0.1	-	-
PEMS-SF	0.05	0.1	0.05	**0.95**	0.25	0.25	-	-
PenDigits	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
PhonemeSpectra	0.1	0.1	0	0.3	0.15	0.05	0	**0.95**
RacketSports	0.3	0.25	0.3	**0.6**	0.25	0.15	0.24	0.55
SelfRegulationSCP1	0.15	0.1	0.1	**1**	**1**	**1**	0.05	**1**
SelfRegulationSCP2	0.05	0.1	0.05	**1**	**1**	**1**	0.05	**1**
SpokenArabicDigits	0.15	0.2	0.2	**1**	0.1	0.05	0.35	**1**
StandWalkJump	0.05	0.35	0.35	0.55	**0.9**	0.85	0.35	0.7
UWaveGestureLibrary	0	0	0.1	**0.2**	0.05	0.05	0.05	0

Table 3: 1 minus the Type II error for sequence length $L_0 = 10$ and $m = n = 30$ samples.

	MMDs			Signature MMDs			Classic Statistics	
	Linear	RBF	Laplace	Linear	RBF	Laplace	Hotelling	Wolf
ArticularyWordRecognition	0	0.1	0.35	**1**	0.65	0.85	0.15	**1**
AtrialFibrillation	0.1	0.25	0.4	**1**	**1**	**1**	0.6	**1**
BasicMotions	0.9	**1**	0.95	**1**	**1**	**1**	0.47	**1**
CharacterTrajectories	0.85	**1**	**1**	**1**	**1**	**1**	**1**	**1**
Cricket	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
DuckDuckGeese	0.2	0.2	0.05	**1**	0.95	0.75	-	-
ERing	0.8	0.95	**1**	**1**	**1**	**1**	0.85	**1**
EigenWorms	0.05	0.3	0.45	**1**	0.35	0.52	0.05	0.5
Epilepsy	0.05	0.15	0.3	0.8	0.4	0.1	0.1	**1**
EthanolConcentration	0.05	0	0.45	0.7	0.05	0.05	0	**1**
FaceDetection	0.05	1	0.65	**0.8**	0.15	0.1	-	-
FingerMovements	**1**	**1**	**1**	**1**	0.25	0.05	-	-
HandMovementDirection	0.05	0	0.1	**1**	**1**	0.6	0.1	**1**
Handwriting	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
Heartbeat	0.1	0	0.15	**0.9**	0.05	0	-	-
InsectWingbeat	**0.2**	0.1	**0.2**	**0.2**	**0.2**	0.15	-	-
JapaneseVowels	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
LSST	0	0.1	0.1	0.45	0	0.05	0.14	**1**
Libras	0.05	0.1	0.15	0.6	0.8	0.55	0.75	**1**
MotorImagery	0.05	0	0.2	**1**	0.7	0.25	-	-
NATOPS	0.15	0.35	0.15	**1**	0.25	0.2	-	-
PEMS-SF	0.4	0.3	0.4	**1**	0.05	0.2	-	-
PenDigits	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
PhonemeSpectra	0	0.15	0	0.7	0.1	0.1	0.05	**1**
RacketSports	0.85	0.8	0.9	0.95	0.55	0.3	0.25	**1**
SelfRegulationSCP1	0.1	0.25	0.1	**1**	**1**	**1**	0.1	**1**
SelfRegulationSCP2	0.05	0.1	0.05	**1**	**1**	**1**	0.05	**1**
SpokenArabicDigits	0.35	0.85	0.65	**1**	0.15	0.2	**1**	**1**
StandWalkJump	0.05	0.6	**1**	**1**	**1**	**1**	0.75	**1**
UWaveGestureLibrary	0.05	0.05	0.1	**0.15**	0.05	0	0.05	0.1

Table 4: 1 minus the Type II error for sequence length $L_0 = 10$ and $m = n = 70$ samples.

	MMDs			Signature MMDs		
	Linear	RBF	Laplace	Linear	RBF	Laplace
ArticularyWordRecognition	**1**	**1**	**1**	**1**	**1**	0.95
AtrialFibrillation	0.45	**0.9**	0.8	0.7	0.7	0.6
BasicMotions	0.5	0.9	0.55	**1**	**1**	**1**
CharacterTrajectories	**1**	**1**	**1**	**1**	**1**	**1**
Cricket	**1**	**1**	**1**	**1**	**1**	**1**
DuckDuckGeese	0.15	0.05	0.2	**0.25**	0.2	0.15
ERing	0.85	**1**	**1**	**1**	**1**	**1**
EigenWorms	0.15	0.65	0.45	**1**	**1**	**1**
Epilepsy	0.05	0.05	0.05	**0.25**	0.1	0.1
EthanolConcentration	0.1	0	0	**0.15**	0.05	0.05
FaceDetection	0	0	0.15	**0.2**	0.05	0.1
FingerMovements	**1**	**1**	**1**	**1**	0.95	**1**
HandMovementDirection	0.6	0.9	0.85	**1**	0.15	0.5
Handwriting	0.75	**1**	**1**	0.9	0.95	0.75
Heartbeat	0.05	0.05	0	**0.4**	0.05	0.1
InsectWingbeat	0	0.05	0.1	0	0.25	0.2
JapaneseVowels	**1**	**1**	**1**	**1**	**1**	**1**
LSST	0.05	0	0.05	0.05	**0.1**	0.05
Libras	0.6	0.4	0.3	**0.9**	0.85	0.5
MotorImagery	0	0	0.1	**1**	**1**	0.6
NATOPS	0.3	0.95	0.95	**1**	0.2	0
PEMS-SF	**1**	0.35	**1**	**1**	**1**	0.15
PenDigits	**1**	**1**	**1**	**1**	0.95	0.9
PhonemeSpectra	0	0	0	**0.3**	0.15	0
RacketSports	0.35	0.65	**0.9**	0.5	0.05	0.25
SelfRegulationSCP1	**1**	**1**	**1**	**1**	**1**	**1**
SelfRegulationSCP2	**1**	**1**	**1**	**1**	**1**	**1**
SpokenArabicDigits	**1**	**1**	**1**	0.95	0	0.1
StandWalkJump	0.1	**0.95**	0.8	0.65	0.75	0.6
UWaveGestureLibrary	**1**	**1**	**1**	0.85	**1**	**1**

Table 5: 1 minus the Type II error for sequence length $L_0 = 100$ and $m = n = 30$ samples.

	MMDs			Signature MMDs		
	Linear	RBF	Laplace	Linear	RBF	Laplace
ArticularyWordRecognition	**1**	**1**	**1**	**1**	**1**	**1**
AtrialFibrillation	**1**	**1**	**1**	**1**	**1**	**1**
BasicMotions	**1**	**1**	**1**	**1**	**1**	**1**
CharacterTrajectories	**1**	**1**	**1**	**1**	**1**	**1**
Cricket	**1**	**1**	**1**	**1**	**1**	**1**
DuckDuckGeese	0.2	0.2	**1**	**1**	**1**	**1**
ERing	**1**	**1**	**1**	**1**	**1**	**1**
EigenWorms	0.45	**1**	**1**	**1**	**1**	**1**
Epilepsy	0.15	0.75	0.3	**0.95**	0.85	0.8
EthanolConcentration	0.15	0.05	0	**0.6**	0.05	0.1
FaceDetection	0.05	0.15	**1**	**1**	**1**	**1**
FingerMovements	**1**	**1**	**1**	**1**	**1**	**1**
HandMovementDirection	**1**	**1**	**1**	**1**	**1**	**1**
Handwriting	**1**	**1**	**1**	**1**	**1**	**1**
Heartbeat	0.1	0.15	0.1	**1**	**1**	**1**
InsectWingbeat	0.1	0.1	0.55	0.05	**1**	**1**
JapaneseVowels	**1**	**1**	**1**	**1**	**1**	**1**
LSST	0.15	**0.85**	0.3	0.45	0.5	0.35
Libras	**1**	**1**	**1**	**1**	**1**	**1**
MotorImagery	0.05	0.2	0.15	**1**	**1**	**1**
NATOPS	**1**	**1**	**1**	**1**	**1**	**1**
PEMS-SF	**1**	**1**	**1**	**1**	**1**	**1**
PenDigits	**1**	**1**	**1**	**1**	**1**	**1**
PhonemeSpectra	0.15	0.05	0.15	**0.7**	0.35	0.4
RacketSports	**1**	**1**	**1**	**1**	**1**	**1**
SelfRegulationSCP1	**1**	**1**	**1**	**1**	**1**	**1**
SelfRegulationSCP2	**1**	**1**	**1**	**1**	**1**	**1**
SpokenArabicDigits	**1**	**1**	**1**	**1**	**1**	**1**
StandWalkJump	**1**	**1**	**1**	**1**	**1**	**1**
UWaveGestureLibrary	**1**	**1**	**1**	**1**	**1**	**1**

Table 6: 1 minus the Type II error for sequence length $L_0 = 200$ and $m = n = 200$ samples.

8. Summary

We generalized the classical moment map from the domain of random variables in finite dimensions to the domain of path-valued random variables. This yields a robust, universal, and characteristic feature map for stochastic processes, that can be kernelized. The associated MMD becomes a metric for laws of stochastic processes that can be efficiently estimated from finite samples. Let us highlight that our non-parametric approach of studying probability measures on $\bigcup_{T>0} C([0,T], \mathcal{X})$ encompasses a large class of examples which are important to different communities, e.g.

- genuinely discrete sequential data such as text,[16]

- classical time series such as (G)ARCH, ARMA, etc. as used in econometrics,

- stochastic differential equations and semimartingales used in stochastic calculus,

- stochastic processes in space and time, e.g. stochastic partial differential equations seen as evolution equations (ODEs with noise in infinite-dimensional state space),

---

16. If $\mathcal{X}$ is linear, sequences in $\mathcal{X}$ embed into paths in $\mathcal{X}$, $\mathcal{X}^n \hookrightarrow C([0,1], \mathcal{X})$, via the Donsker embedding $(x_i)_{i=1}^n \mapsto \left(t \mapsto x_{\lfloor nt \rfloor} + (nt - \lfloor nt \rfloor) \sum_{i=1}^{\lceil nt \rceil} x_i \right)$. This pushes measures on sequences to measures on paths and allows to treat genuinely discrete data in our framework; e.g. for text, $\mathcal{X}$ is the free vector space spanned by letters and the path associated with a text (a sequence of letters) is a lattice path; multi-variate time series are sequences in $\mathcal{X} = \mathbb{R}^d$, etc.

- persistence paths as they arise in topological data analysis (Chevyrev et al., 2020),

- the evolution of structured, non-Euclidean objects, such as networks, molecules, images, provided a kernel for the state space $\mathcal{X}$ in which they evolve is given.

From a theoretical perspective, the existence of a universal and characteristic feature map for stochastic processes opens several research venues; two of them are

Statistics in high/infinite dimensions Non-parametric testing in high dimensions suffers from decreasing test power (Ramdas et al., 2015). Indeed, the classic kernels perform poorly in our experiments, but the robust signature kernel does well. This is not too surprising from a stochastic analysis perspective since sequences of iterated integrals provide a natural basis for functions of paths. Signatures are classical in stochastic analysis, but applications in statistics are much more recent, e.g. Papavasiliou and Ladroue (2011) use it for SDE parameter estimation and it would be interesting to connect this with the methods developed in this paper.

Parametric Statistics The focus of this article is the non-parametric learning of laws of stochastic processes. However, for some applications it can be useful to focus on a small class of stochastic processes, e.g. certain parameterized stochastic differential equations as they appear in finance. For some of these cases, semi-explicit formulas for the resulting signature moments are known (as for Brownian motion in Example 2 or more generally Lévy processes (Friz and Shekhar, 2017; Chevyrev, 2018)), or can be described by a PDE, see (Lyons and Ni, 2015). It is natural to ask whether such strong model assumptions can be leveraged to yield signature MMDs with (semi-)explicit expressions.

Acknowledgments

A. Tensors and the Tensor Algebra $\prod_{m \geq 0} V^{\otimes m}$

For general background on tensors we refer to (Lang, 2002, Chapter 16). Recall that if V, W are vector spaces (possibly infinite-dimensional) then there exists another vector space $V \otimes W$ and a bilinear map $\varphi : V \times W \to V \otimes W$ with the universal property that any other bilinear map $V \times W \to Z$ factors uniquely through φ.[17] This map $\otimes$ is called the tensor product and we call the elements of $V \otimes W$ tensors. In particular, for $m \geq 0$, we call an element of $V \otimes \cdots \otimes V = V^{\otimes m}$ a tensor of degree m; by convention, $V^{\otimes 0} = \mathbb{R}$. For any vector space V, the space

$$\prod_{m \geq 0} V^{\otimes m} = \{\mathbf{t} = (\mathbf{t}^m)_{m \geq 0} \,|\, \mathbf{t}^m \in V^{\otimes m}\}$$

---

17. Concretely, if $A \subset V$ and $B \subset W$ are bases of V and W respectively, then $\{a \otimes b : a \in A, \ b \in B\}$ is a basis of $V \otimes W$.

is also a vector space but also carries a product. This product is the so-called *tensor convolution product* and is defined by[18]

$$\mathbf{s} \otimes \mathbf{t} := \Big(\sum_{i=0}^{m} \mathbf{s}^i \otimes \mathbf{t}^{m-i} \Big)_{m \geq 0} = (\mathbf{s}^0 \mathbf{t}^0, \mathbf{s}^1 \mathbf{t}^0 + \mathbf{s}^0 \mathbf{t}^1, \mathbf{s}^2 \mathbf{t}^0 + \mathbf{s}^1 \otimes \mathbf{t}^1 + \mathbf{s}^0 \mathbf{t}^2, \dots). \tag{20}$$

The unit element for this product is $\mathbf{1} := (1, 0, 0, \dots)$.

The reason why $\prod_{m \geq 0} V^{\otimes m}$ is a fundamental mathematical object is that $\prod_{m \geq 0} V^{\otimes m}$ is the most general algebra containing V, namely $\prod_{m \geq 0} V^{\otimes m}$ *is the "free algebra" that contains* V, see (Reutenauer, 1993, Chapter 1). Further, it is convenient to define

$$\exp : \prod_{m \geq 0} V^{\otimes m} \to \prod_{m \geq 0} V^{\otimes m}, \quad \exp(\mathbf{s}) := \sum_{m \geq 0} \frac{\mathbf{s}^{\otimes m}}{m!}.$$

(Applied with $\mathbf{s} = (0, v, 0, 0, \dots)$ for $v \in \mathbb{R}^d$ this reduces to the usual vector exponential). In the main text we also use the truncated tensor algebra $\prod_{m=0}^{M} V^{\otimes m}$ which is again a linear space and also forms an algebra with the tensor convolution product (20) restricted to the first M tensors.

B. Features for Geometric Rough Paths

In this appendix we give a brief introduction to rough paths theory. For a detailed introduction see the lecture notes (Lyons et al., 2007) or the monographs (Friz and Hairer, 2014; Friz and Victoir, 2010; Lyons and Qian, 2002).

To define the iterated integrals $\int dx^{\otimes m}$ using Riemann–Stieltjes–Young integration requires $x : [0, T] \to V$ to have finite p-variation for some $p \in [1, 2)$. In particular, many processes of interest in stochastic analysis fall outside this scope. The rough paths approach is to flesh out the abstract properties of maps that associate with a path x over a time interval $[s, t]$, an element of $\prod_{m=0}^{\lfloor p \rfloor} V^{\otimes m}$ such that it "behaves like" $\int dx^{\otimes m}$.

Throughout this appendix, let V be a Banach space and $p \geq 1$. Recall from Appendix A that $\prod_{m=0}^{M} V^{\otimes m}$ is an algebra with tensor multiplication (20).

Definition 34 *A p-rough path is a continuous map* $\mathbf{x} = (\mathbf{x}^0, \mathbf{x}^1, \dots, \mathbf{x}^{\lfloor p \rfloor}) : [0, T_\mathbf{x}] \to \prod_{m=0}^{\lfloor p \rfloor} V^{\otimes m}$ *such that* $\mathbf{x}^0 \equiv 1$ *and, denoting[19]* $\mathbf{x}(s, t) := \mathbf{x}(s)^{-1} \otimes \mathbf{x}(t)$, *for all* $0 \leq m \leq \lfloor p \rfloor$

$$\sup_{\substack{n \geq 1 \\ 0 \leq t_1 \leq \dots \leq t_n \leq T_\mathbf{x}}} \sum_{i=1}^{n-1} \|\mathbf{x}^m(t_i, t_{i+1})\|^{p/m} < \infty. \tag{21}$$

If $x \in C([0, T_x], V)$ and there exists a p-rough path $\mathbf{x}$ such that $\mathbf{x}^1(t) = x(t)$, then we call $\mathbf{x}$ a p-rough path lift of x.

---

18. Since $V^{\otimes 0} = \mathbb{R}$, note that $\mathbf{s}^0, \mathbf{t}^0$ are real numbers, and therefore so is $\mathbf{s}^0 \mathbf{t}^0 \in V^{\otimes 0} = \mathbb{R}$. Furthermore, $\mathbf{s}^1 \mathbf{t}^0, \mathbf{s}^0 \mathbf{t}^1, \mathbf{s}^2 \mathbf{t}^0, \mathbf{s}^0 \mathbf{t}^2, \dots$ are well-defined products of a tensor with a scalar.

19. Here, for $\mathbf{t} \in \prod_{m=0}^{\lfloor p \rfloor} V^{\otimes m}$ with $\mathbf{t}^0 = 1$, we denote by $\mathbf{t}^{-1}$ the unique element of $\prod_{m=0}^{\lfloor p \rfloor} V^{\otimes m}$ such that $\mathbf{t} \otimes \mathbf{t}^{-1} = \mathbf{1}$ (the set of all $\mathbf{t} \in \prod_{m=0}^{\lfloor p \rfloor} V^{\otimes m}$ with $\mathbf{t}^0 = 1$ is readily seen to be a group, so in particular $\mathbf{t}^{-1}$ exists).

For $p = 1$, we recover the setting of Section 4 since every $x \in \mathcal{C}^1$ is a 1-rough path by identifying $\prod_{m=0}^{1} V^{\otimes m} = \mathbb{R} \oplus V$ with V. Furthermore, for any $p \geq 1$, every $x \in \mathcal{C}^1$ canonically defines a p-rough path $\mathbf{x}$ by setting $\mathbf{x}^m(t) = \int_0^t dx^{\otimes m}$, where we used the notation (12). We call $\mathbf{x}$ the canonical p-rough path lift of x.

A p-rough path $\mathbf{x}$ is called *geometric* if there exist p-rough paths $(\mathbf{x}_n)_{n \geq 1}$ which are the canonical lifts of smooth paths such that $\mathbf{x}_n \to \mathbf{x}$ in the p-variation metric $d_{p-\text{var}}$.[20]

Example 8 *(1) (Semi)martingales. With $2 < p < 3$ we cover the theory of continuous (semi)martingales (Friz and Victoir, 2010, Sec. 14): every continuous semimartingale $X : [0, T] \to \mathbb{R}^d$ is of bounded p-variation for any $p > 2$. Stochastic (Itô or Stratonovich) integration can be used to the define the first $2 = \lfloor p \rfloor$ iterated integrals $\mathbf{X}(t) = (1, \int_0^t dX, \int_0^t dX^{\otimes 2})$. One can verify that (21) holds almost surely, hence $\mathbf{X}$ is a p-rough path. Moreover, if Stratonovich integration is used, then $\mathbf{X}$ is geometric. Extensions exist to the discontinuous setting (Chevyrev and Friz, 2019).*

(2) Gaussian processes. Many Gaussian processes admit canonical lifts to geometric p-rough paths. A criterion for such a lift to exist can be given in terms of the covariance function (Coutin and Qian, 2002; Friz et al., 2016), which in particular covers fractional Brownian motion with Hurst parameter $H > 1/4$.

(3) Markov processes. Likewise, many Markov processes admit canonical lifts to geometric p-rough paths (typically $2 < p < 3$). These include diffusions on fractals (Bass et al., 2002) and Markov processes arising from elliptic Dirichlet forms (Friz and Victoir, 2008; Chevyrev and Ogrodnik, 2018). Such processes are typically not semimartingales and thus fall outside the scope of Itô–Stratonovich calculus.

To mimic the construction in Section 4, we define the space of geometric p-rough paths

$$\mathcal{C}^p := \bigcup_{T > 0} \left\{ \mathbf{x} : [0, T] \to \prod_{m=0}^{\lfloor p \rfloor} V^{\otimes m} : \mathbf{x} \text{ is a geometric } p\text{-rough path} \right\},$$

equipped with the (non-Hausdorff) topology induced by

$$\rho_p(\mathbf{x}, \mathbf{y}) = \inf_{\tau} d_{p-\text{var}}(\mathbf{x}, \mathbf{y} \circ \tau), \tag{22}$$

where the infimum is over all increasing bijections $\tau : [0, T_{\mathbf{x}}] \to [0, T_{\mathbf{y}}]$.

Remark 35 *In the case $V = \mathbb{R}^d$, a geometric p-rough path satisfies the shuffle identity (14).*

B.1 Tree-like Equivalence

For a topological space $\mathcal{X}$, a function $x : [0, T_x] \to \mathcal{X}$ is called *tree-like* if x is continuous and there exists an $\mathbb{R}$-tree $\mathfrak{T}$, a continuous function $\phi : [0, T_x] \to \mathfrak{T}$, and a map $\psi : \mathfrak{T} \to \mathcal{X}$ such that

---

20. See (Lyons and Qian, 2002, Eq. (3.70)) for the definition of the p-variation metric; for brevity, we do not give a definition here, particularly because its explicit form plays no role in the sequel.

$\phi(0) = \phi(T_x)$ and $x = \psi \circ \phi$. Let $\overleftarrow{x} : [0, T_x] \to \mathcal{X}$, $\overleftarrow{x}(t) = x(T_x - t)$ denote the time-reversal of x. For another function $y : [0, T_y] \to \mathcal{X}$, we denote the concatenation of x with y by

$$x * y : [0, T_x + T_y] \to \mathcal{X} , \quad x * y(t) = \begin{cases} x(t) & \text{if } t \in [0, T_x] \\ y(t) & \text{if } t \in (T_x, T_x + T_y] . \end{cases}$$

Note that $x * y$ is a continuous path if x and y are continuous and $x(T_x) = y(0)$. We say that x and y are *tree-like equivalent*, and write $x \sim_t y$, if $x * \overleftarrow{y}$ is tree-like.

Example 9 *An important example of tree-like equivalence is when x is a reparameterization of y i.e. there exists an increasing bijection $\tau : [0, T_x] \to [0, T_y]$ such that $x = y \circ \tau$.*

Example 10 *Another way in which paths can be tree-like equivalent is if they differ only through back-tracking excursions. For example the following two paths $x, y : [0, 4] \to \mathbb{R}^2$ depicted in the figures below are tree-like equivalent: $x(t) := [t, t]$ and*

$$y(t) := \begin{cases} [2t, 2t] & \text{if } t \in [0, 1) \\ [2 + 2(t-1), 2 - 2(t-1)] & \text{if } t \in [1, 2) \\ [4 - 2(t-2), 2(t-2)] & \text{if } t \in [2, 3) \\ [2 + 2(t-3), 2 + 2(t-3)] & \text{if } t \in [3, 4] . \end{cases}$$
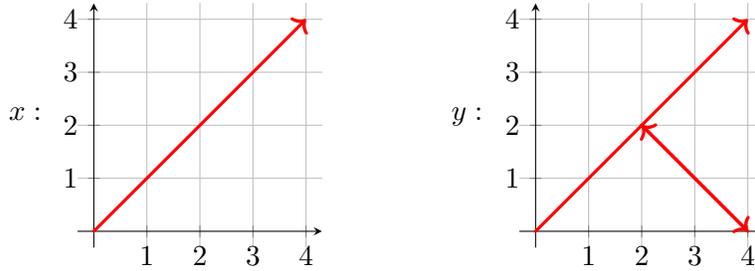
B.2 Ordered Moments for Rough Paths

A key result of Lyons (1998) is that the signature map S defined by (13) extends to the space of p-rough paths: for any $\mathbf{x} \in \mathcal{C}^p$, the integrals $\int d\mathbf{x}^{\otimes m}$ are canonically defined for all $m > \lfloor p \rfloor$. Moreover, S is injective on $\mathcal{C}^p$ up to tree-like equivalence (Boedihardjo et al., 2016). We summarise these results in the following extension of Theorem 19.

Theorem 36 *There exists a map $S : \mathcal{C}^p \to \mathbf{T}_1(V)$, $S(\mathbf{x}) = (1, S^1(\mathbf{x}), S^2(\mathbf{x}), \ldots)$ such that $S^m(\mathbf{x}) = \mathbf{x}^m(T_\mathbf{x})$ for all $m = 0, \ldots, \lfloor p \rfloor$. Furthermore, $S(\mathbf{x}) = S(\mathbf{y})$ if and only if $\mathbf{x} \sim_t \mathbf{y}$.*

Example 11

- *Let $x \in \mathcal{C}^1$. Then $S(x) \in \mathbf{T}_1(V)$ defined by (13) agrees with $S(x)$ in Theorem 36.*

- *Let $X = (X(t))_{t \in [0, T]}$ be a continuous semimartingale in $\mathbb{R}^d$ with geometric p-rough path lift $\mathbf{X} = (1, \mathbf{X}^1, \mathbf{X}^2) \equiv (1, \int dX, \int dX^{\otimes 2})$ as in part (1) of Example 8. Then*

$$S(\mathbf{X}) = \left(1, \int_0^T dX, \int_0^T dX^{\otimes 2}, \int_0^T dX^{\otimes 3}, \ldots \right) \in \mathbf{T}_1(\mathbb{R}^d),$$

where the stochastic integrals are taken in the Stratonovich sense.

Theorem 36 suggests the following generalization of Definition 20.

Definition 37 *We define the space of unparameterized geometric p-rough paths $\mathcal{P}^p$ as the set of equivalence classes $\mathcal{C}^p / \sim_t$. We equip $\mathcal{P}^p$ with the topology induced by the signature map embedding* $\mathrm{S} : \mathcal{P}^p \to \mathbf{T}_1(V)$.

B.3 Time Parameterization

As in the bounded variation setting of Section 4.3, we can make the map S variant to the parameterization of time by adding a time component. For any $\mathbf{x} \in \mathcal{C}^p(V)$, there is a canonical p-rough path lift of $t \mapsto (\mathbf{x}^1(t), t)$, denoted by $\overline{\mathbf{x}} \in \mathcal{C}^p(V \oplus \mathbb{R})$, and which extends $\mathbf{x}$ in the sense that $\langle \ell, \mathbf{x}^m \rangle = \langle \ell, \overline{\mathbf{x}}^m \rangle$ for all $m = 0, \ldots, \lfloor p \rfloor$ and $\ell \in (V^{\otimes m})'$, see (Lyons and Qian, 2002, Sec. 3.3.3). As such, we can identify $\mathcal{C}^p(V)$ with a subset of $\mathcal{C}^p(V \oplus \mathbb{R})$. Furthermore, $\overline{\mathbf{x}} \sim_t \overline{\mathbf{y}}$ if and only if $\mathbf{x} = \mathbf{y}$, and thus $(\mathcal{C}^p(V), \overline{\rho})$ is a genuine metric space, where $\overline{\rho}_p$ is the quantity (22) associated to $\mathcal{C}^p(V \oplus \mathbb{R})$.

B.4 Topological Properties and MMD for Measures on Rough Paths

Lemma 38 *Let $(\mathbf{x}_n)_{n=1}^\infty$ be a Cauchy sequence for ρ_p. Then $\rho_p(\mathbf{x}_n, \mathbf{x}) \to 0$ for some $\mathbf{x} \in \mathcal{C}^p$.*

Proof Denoting by $\mathcal{C}^p_{[0,T]}$ the space of geometric rough paths defined on $[0, T]$, recall that $(\mathcal{C}^p_{[0,T]}, d_{p-\mathrm{var}})$ is a complete metric space (Lyons and Qian, 2002, Lem. 3.3.3). We can find a subsequence $n(k)$ and time-changes $\tau_k : [0, 1] \to [0, T_{\mathbf{x}_{n(k)}}]$ such that, denoting $\mathbf{y}_k = \mathbf{x}_{n(k)} \circ \tau_k \in \mathcal{C}^p_{[0,1]}$, $d_{p-\mathrm{var}}(\mathbf{y}_k, \mathbf{y}_{k+1}) < 2^{-k}$. In particular, $\mathbf{y}_k$ is Cauchy for $d_{p-\mathrm{var}}$ and converges to some $\mathbf{x} \in \mathcal{C}^p_{[0,1]}$, which implies that $\rho_p(\mathbf{x}_n, \mathbf{x}) \to 0$. ∎

Proposition 39 *The topological space $\mathcal{P}^p$ is metrizable and the metric space $(\mathcal{C}^p, \overline{\rho}_p)$ is complete. Moreover, if V is separable, then so are $(\mathcal{C}^p, \rho_p)$ and $(\mathcal{C}^p, \overline{\rho}_p)$.*

Proof Since S is injective and $\mathbf{T}_1(V)$ is a metric space, $\mathcal{P}^p$ is metrizable. Completeness of $(\mathcal{C}^p, \overline{\rho}_p)$ follows from Lemma 38 and the fact that $\{\overline{\mathbf{x}} : \mathbf{x} \in \mathcal{C}^p\}$ is a closed subset of $(\mathcal{C}^p(V \oplus \mathbb{R}, \overline{\rho}_p)$. If V is separable, then $C^\infty([0, T], V)$ is separable under the C^1-norm $\| \cdot \|_{C^1}$, see (Friz and Hairer, 2014, Sol. 2.13). Since the canonical lift map $(C^\infty, \| \cdot \|_{C^1}) \ni x \mapsto \mathbf{x} \in (\mathcal{C}^p, \rho_p)$ is continuous, separability of $(\mathcal{C}^p, \rho_p)$ and $(\mathcal{C}^p, \overline{\rho}_p)$ follows. ∎

We conclude this section with the proof of Theorem 30(iv). Let notation be as in Section 6.2 and Theorem 30.

Proposition 40   *(i) There exist probability measures μ_n, μ, on $\mathcal{P}^p$ such that $d_\mathrm{k}(\mu_n, \mu) \to 0$ but such that μ_n does not converge weakly to μ.*

  (ii) Suppose H is finite dimensional and that $\mathcal{M}$ is a set of probability measures on $\mathcal{P}^p$ which is compact under weak convergence. Then d_k and weak convergence induce the same topology on $\mathcal{M}$. In particular, weak convergence implies convergence in d_k.
  The same statements hold with $\mathcal{P}^p$ replaced by $(\mathcal{C}^p, \overline{\rho}_p)$.

Proof (i) Let $p < p' < \lfloor p \rfloor + 1$. Consider $\mathbf{x} \in \mathcal{P}^p$ and a sequence $\mathbf{x}_n \in \mathcal{P}^p$ such that $\mathbf{x}_n$ does not converge to $\mathbf{x}$ as elements of of $\mathcal{P}^p$ but such that $\mathbf{x}_n \to \mathbf{x}$ as elements of $\mathcal{P}^{p'}$, i.e., in the p'-variation metric (it is a simple exercise to construct such a sequence for any $\mathbf{x} \in \mathcal{P}^p$). Due to continuity of $\Phi : \mathcal{P}^{p'} \to \mathbf{T}(H)$, we conclude that $\|\Phi(\mathbf{x}_n) - \Phi(\mathbf{x})\| \to 0$. In particular, for the corresponding Dirac delta measures, we have

$$d_{\mathrm{k}}(\delta_{\mathbf{x}_n}, \delta_{\mathbf{x}}) = \sup_{\mathbf{t} \in \mathbf{T}(H), \|\mathbf{t}\| \leq 1} |\langle \mathbf{t}, \Phi(\mathbf{x}_n) \rangle| - |\langle \mathbf{t}, \Phi(\mathbf{x}) \rangle| \to 0.$$

However, due to the assumption that $\mathbf{x}_n$ does not converge to $\mathbf{x}$ in $\mathcal{P}^p$, it holds that $\delta_{\mathbf{x}_n}$ does not converge weakly to $\delta_{\mathbf{x}}$ as probability measures on $\mathcal{P}^p$.

(ii) Let μ and μ_n be probability measures on $\mathcal{P}^p$ such that $\mu_n \to \mu$ weakly. Observe that

$$
\begin{aligned}
d_{\mathrm{k}}(\mu_n, \mu)^2 &= \sup_{\mathbf{t} \in \mathbf{T}(H), \|\mathbf{t}\| \leq 1} \left| \int_{\mathcal{P}^p} \langle \mathbf{t}, \Phi(\mathbf{x}) \rangle \mu_n(d\mathbf{x}) - \int_{\mathcal{P}^p} \langle \mathbf{t}, \Phi(\mathbf{x}) \rangle \mu(d\mathbf{x}) \right|^2 \\
&= \left\| \int_{\mathcal{P}^p} \Phi(\mathbf{x}) \mu_n(d\mathbf{x}) - \int_{\mathcal{P}^p} \Phi(\mathbf{x}) \mu(d\mathbf{x}) \right\|^2_{\mathbf{T}(H)} \\
&= \sum_{m \geq 0} \left\| \int_{\mathcal{P}^p} \lambda(\mathrm{S}(\mathbf{x}))^m \, \mathrm{S}^m(\mathbf{x}) \mu_n(\mathbf{x}) - \int_{\mathcal{P}^p} \lambda(\mathrm{S}(\mathbf{x}))^m \, \mathrm{S}^m(\mathbf{x}) \mu(\mathbf{x}) \right\|^2_{H^{\otimes m}}.
\end{aligned}
$$

Note that for every $m \geq 0$, $\mathbf{x} \mapsto \lambda(\mathrm{S}(\mathbf{x}))^m \, \mathrm{S}^m(\mathbf{x})$ is a continuous bounded function from $\mathcal{P}^p$ into the finite dimensional vector space $H^{\otimes m}$. It follows from the weak convergence $\mu_n \to \mu$ that the final quantity converges to zero as $n \to \infty$, thus weak convergence implies convergence in d_{k}. Therefore $\mathcal{M}$ is also compact under d_{k}, and since the topologies induced by d_{k} and weak convergence are comparable, the two topologies necessarily coincide on $\mathcal{M}$. Therefore $\mathcal{M}$ is also compact under d_{k}, and since the topologies induced by d_{k} and weak convergence are comparable, the two topologies necessarily coincide on $\mathcal{M}$.

The corresponding claims for $\mathcal{C}^p$ follow in an identical manner. ∎

B.5 Example: Non-robust Signature Moments gone wrong

Define $X, Y : [0, 1] \to \mathbb{R}^2$ by $X_t = t \cdot N^\top$ and $Y_t = t \cdot M^\top$, where $N = (N_1, N_2)$ consists of two-independent lognormal distributions and $M = (M_1, M_2)$ consists of two independent perturbed lognormals, i.e. $p(n_1, n_2) = \prod_{i=1}^n \frac{1}{n_i \sqrt{2\pi}} \exp(\frac{-\log^2(n_i)}{2})$ and $q(m_1, m_2) = p(m_1, m_2) \prod_{i=1}^2 (1 + \sin(2\pi \log m_i))$. Then all signature moments of X and Y coincide although X and Y have different laws,

$$\mathbb{E}[\int dX^{\otimes m}] = \mathbb{E}[\int dY^{\otimes m}] \text{ for all } m \geq 0 . \tag{23}$$

The moment equality (23) holds since a direct calculation shows that $\int dX^{\otimes m} = \frac{(X(1) - X(0))^{\otimes m}}{m!} = \frac{N^{\otimes m}}{m!}$ and analogously $\int dY^{\otimes m} = \frac{M^{\otimes m}}{m!}$. Hence the only information captured in $\mathbb{E}[\int dX^{\otimes m}]$ and $\mathbb{E}[\int dY^{\otimes m}]$ are the m-th moments of M and N, but $\mathbb{E}[M^{\otimes m}] = \mathbb{E}[N^{\otimes m}]$ follows from a direct calculation.

C. Kernel Background

Throughout this appendix, we fix a topological space $\mathcal{X}$, an inner product space $(E, \langle \cdot, \cdot \rangle_E)$ and feature map $\Phi : \mathcal{X} \to E$. For $x \in \mathcal{X}$, we define the kernel function $k_x \in \mathbb{R}^{\mathcal{X}}$ by $k_x(y) := k(x, y)$, where $k(x, y) := \langle \Phi(x), \Phi(y) \rangle$. Consider the subspace $\mathcal{H}_0 := \mathrm{span}\{k_x : x \in \mathcal{X}\} \subset \mathbb{R}^{\mathcal{X}}$ equipped with the inner product defined by $\langle k_x, k_y \rangle_{\mathcal{H}_0} := k(x, y)$. We first recall the following theorem of Moore–Aronszajn.

Theorem 41 (Moore–Aronszajn) *There exits a unique Hilbert space $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ with k as the reproducing kernel. Moreover, $\mathcal{H}_0$ is dense in $\mathcal{H}$.*

The following theorem clarifies the relation to our original feature map $x \mapsto \Phi(x)$ and $x \mapsto k_x$ and shows how to construct $\mathcal{H}$ as a subspace of (the completion of) E.

Theorem 42 *(i) There exists a unique linear map $\Psi : \mathcal{H}_0 \to E$ such that the following diagram commutes:*

$$
\begin{array}{ccc}
\mathcal{X} & \xrightarrow{x \to k_x} & \mathcal{H}_0 \\
& {\scriptstyle \Phi} \searrow & \downarrow {\scriptstyle \Psi} \\
& & E \ .
\end{array}
$$

(ii) The map Ψ is injective and is an isometry onto its image.

(iii) Denote by $x \mapsto x'$ the canonical map $E \to E'$ which identifies E with (a subspace of) E'. Then

$$\imath(\Phi(x)') = k_x$$

and the following diagram commutes:

$$
\begin{array}{ccccc}
\mathcal{X} & \xrightarrow{x \mapsto k_x} & \mathcal{H}_0 & \xhookrightarrow{\mathrm{id}} & \mathbb{R}^{\mathcal{X}} \\
& {\scriptstyle \Phi} \searrow & \downarrow {\scriptstyle \Psi} & & \uparrow {\scriptstyle \imath} \\
& & E & \xrightarrow{x \mapsto x'} & E' \ .
\end{array}
$$

(iv) The image of $\mathcal{H}_0$ in E' under $h \mapsto \Psi(h)'$ is a dense subspace of $\mathrm{Ker}(\imath)^{\perp}$.

Proof For Point (i), existence and uniqueness of Ψ follows from the observation that if $\sum a_i k_{x_i} \equiv 0$, then $\sum a_i \langle \Phi(x_i), \Phi(y) \rangle = 0$ for all $y \in \mathcal{X}$, and thus $\sum a_i \Phi(x_i)$ is an element of both $\Phi(\mathcal{X})^{\perp}$ and $\mathrm{span}\,[\Phi(\mathcal{X})]$, and thus must be zero. For Point (ii), note that $\Psi(\sum a_i k_{x_i}) = 0$ is equivalent to $\sum a_i \Phi(x_i) = 0$, so that

$$\sum a_i \langle \Phi(x_i), \Phi(y) \rangle = \sum a_i k_{x_i}(y) = 0, \quad \forall y \in \mathcal{X}.$$

It follows that Ψ is injective. The fact that Ψ is an isometry follows from Point (i). Point (iii) follows immediately since for all $x, y \in \mathcal{X}$

$$\imath(\Phi(x)')(y) = \Phi(x)' \circ \Phi(y) = \langle \Phi(x), \Phi(y) \rangle = k(x, y) = k_x(y).$$

For Point (iv), we define for any subset $F \subset E$ the set $F^{\circ} := \{f' \in E' \mid f'(f) = 0, \forall f \in F\}$. Now $\mathrm{Ker}(\imath)$ consists of all $z \in E'$ such that $z'(\Phi(x)) = 0$ for all $x \in \mathcal{X}$, so that $\mathrm{Ker}(\imath) =$

$\Phi(\mathcal{X})^\circ \subset E'$. Since span $[\Phi(\mathcal{X})] = \Psi(\mathcal{H}_0)$, it follows that $\mathrm{Ker}(\imath) = \Psi(\mathcal{H}_0)^\circ$. The conclusion now follows from the fact that for any subspace $F \subset E$, the image of F under $x \mapsto x'$ is dense in $(F^\circ)^\perp$. ∎

Proposition 43 *Suppose $\mathcal{F} \subset \mathbb{R}^\mathcal{X}$ is a locally convex TVS and that $\mathcal{H}_0$ continuously embeds into $\mathcal{F}$. Then the map $\imath : E' \to \mathbb{R}^\mathcal{X}$ given by (9) maps E' continuously into $\mathcal{F}$. Furthermore*

- *Φ is universal to $\mathcal{F}$ iff the kernel k is universal to $\mathcal{F}$,*

- *Φ is characteristic to $\mathcal{F}'$ iff the kernel k is characteristic to $\mathcal{F}'$.*

Proof Substituting $\mathcal{F}$ by its completion if necessary, we may assume $\mathcal{F}$ is complete. Write $F_0 := \Psi(\mathcal{H}_0)' \subset E'$ and let $F \subset E'$ denote the closure of F_0 in E'. By Point (iii) of Theorem 42, it holds that

$$\mathrm{id}(\mathcal{H}_0) = \imath(F_0) \subset \mathcal{F}, \tag{24}$$

so by the assumption that $\mathrm{id} : \mathcal{H}_0 \hookrightarrow \mathcal{F}$ is continuous, the restriction $\imath|_{F_0} : F_0 \to \mathcal{F}$ is continuous. By definition of $\imath$, it is easy to see that the restriction $\imath|_F$ agrees with the unique continuous extension of $\imath|_{F_0}$ to F. Hence $\imath|_F : F \to \mathcal{F}$ is continuous. We now write $E' = F \oplus F^\perp$. By Point (iv) of Theorem 42, we have $F = \mathrm{Ker}(\imath)^\perp$.

Note that $\mathrm{Ker}(\imath)$ is closed in the weak topology of E', and thus, *a fortiori*, under the strong (norm) topology. Indeed, if $\ell_n \to \ell$ pointwise in E', then for all $x \in \mathcal{X}$, $\lim_{n \to \infty} \ell_n(\Phi(x)) = \ell(\Phi(x))$. In particular, if $\ell_n \in \mathrm{Ker}(\imath)$ for all $n \geq 1$, then $\ell \in \mathrm{Ker}(\imath)$.

As a consequence, it holds that $\mathrm{Ker}(\imath) = F^\perp$, and thus

$$E' = F \oplus \mathrm{Ker}(\imath), \tag{25}$$

from which the continuity of $\imath : E' \to \mathcal{F}$ follows.

Finally, by (25) and the continuity of $\imath$, the closures in $\mathcal{F}$ of $\imath(F_0)$ and $\imath(E')$ coincide. Combining with (24), it holds that $\imath(E')$ is dense in $\mathcal{F}$ iff $\mathrm{id}(\mathcal{H}_0)$ is dense in $\mathcal{F}$, from which the first equivalence follows. The second equivalence now follows from Theorem 7 and Proposition 28. ∎

C.1 Signature Kernel Discretization

Suppose we are in the setting of Section 6.4.

Proposition 44 *(i) The kernel $\mathrm{k}_M^{\kappa,+} : \mathcal{X}^+ \times \mathcal{X}^+ \to \mathbb{R}$ is a bounded, positive semidefinite kernel for $\mathcal{X}^+$.*

Moreover, for $x, y \in \mathcal{C}^1(\mathcal{X})$ and partitions $\pi \subset [0, T_x], \pi' \subset [0, T_y]$

$$\left| \mathrm{k}^\kappa(x, y) - \mathrm{k}_M^{\kappa,+}(x^\pi, y^{\pi'}) \right| \leq \overline{K}(F(Q_x) + F(Q_y) + F(\Delta_x^\pi) + F(\Delta_y^{\pi'})), \tag{26}$$

where x^π (resp. $y^{\pi'}$) denote the sequences given by sampling x, y along π (resp. π') and

$$\Delta_x^\pi := \|\kappa_x\|_{1-\mathrm{var}} e^{\|\kappa_x\|_{1-\mathrm{var}}} \max_{t_i \in \pi} \|\kappa_x\|_{1-\mathrm{var};[t_{i-1},t_i]} ,$$

$$Q_x := \frac{e^{\|\kappa_x\|_{1-\mathrm{var}}} \|\kappa_x\|_{1-\mathrm{var}}^{M+1}}{(M+1)!}$$

$$\overline{K} := \|\psi\|_\infty^{1/2}(1 + K^{1/2} + \sqrt{2}\|\psi\|_\infty^{1/2}) ,$$

$$F(x) := x \vee \sqrt{x} .$$

(ii) Consider probability measures μ,ν on $\mathcal{C}^1(\mathcal{X})$ whose marginals on $\mathcal{X}^+$ along partitions π,π' are $\mu^\pi,\nu^{\pi'}$. Then

$$|d_{\mathrm{k}^\kappa}(\mu,\nu) - d_{\mathrm{k}_M^{\kappa,+}}(\mu^\pi,\nu^{\pi'})| \le 4\overline{K}\mathbb{E}[F(Q_X) + F(\Delta_X^\pi) + F(Q_Y) + F(\Delta_Y^{\pi'})] .$$

The same inequality applies if the partitions are also random (see the proof for the formal statement).

Proof For brevity we denote $\mathrm{S}_M^{\mathrm{linear}}(h) := \prod_{i=1}^\ell (1+h_i-h_{i-1})$. (Kiraly and Oberhauser, 2019, Cor. 4.3) implies that

$$\| \mathrm{S}_M^{\mathrm{linear}}(\kappa_{x^\pi}) - \mathrm{S}_M(\kappa_x)\| \le \Delta_x^\pi. \tag{27}$$

The first part of (i) simply follows from the boundedness of Λ and the fact that an inner product, in this case on $\prod_{m=0}^M H^{\otimes m}$, is positive definite. The remaining estimate (26) follows by using (27) in the technical Lemmas 45 and 46 that quantifies the factorial decay in the truncation level M.

For point (ii) we directly prove the more general statement that allows for random partitions. Therefore denote $\Pi := \cup_{T>0}\Pi(T)$ with $\Pi(T) := \{\pi : \pi = \{0 \le t_1 < \cdots \le T\}\}$, and set $\mathcal{S} := \{(x,\pi) \in \mathcal{C}^1(\mathcal{X}) \times \Pi : \pi \in \Pi(T_x)\}$. We define a kernel on $\mathcal{S}$ as $\mathrm{k}_M^{\kappa,\Pi}((x,\pi),(y,\pi')) := \mathrm{k}_M^{\kappa,+}(x^\pi, y^{\pi'})$. Note that the associated MMD $d_{\mathrm{k}_M^{\kappa,\Pi}}$ equals $d_{\mathrm{k}_M^{\kappa,+}}$ for the special case of deterministic partitions. The claim then follows by using the representation of MMD as a sum of expectations of kernels (equation (15)), and point (i). ∎

In the proof of Proposition 44, we used the following two lemmas. We keep notation as in Proposition 44.

Lemma 45 *For any $\mathbf{s},\bar{\mathbf{s}},\mathbf{t},\bar{\mathbf{t}} \in \mathbf{T}_1(H)$*

$$|\langle \Lambda(\mathbf{s}), \Lambda(\mathbf{t})\rangle - \langle \Lambda(\bar{\mathbf{s}}), \Lambda(\bar{\mathbf{t}})\rangle| \le \overline{K}(F(\|\mathbf{s} - \bar{\mathbf{s}}\|) + F(\|\mathbf{t} - \bar{\mathbf{t}}\|)) . \tag{28}$$

Proof This follows from Proposition 14(i)-(iii) and the Cauchy–Schwarz inequality. ∎

Setting $\varphi(x) = \kappa(x,\cdot)$ as in Proposition 31, recall the space $\mathcal{C}^1(\mathcal{X})$ from Definition 24.

Lemma 46 *Let $x,y \in \mathcal{C}^1(\mathcal{X})$. Then*

$$|\mathrm{k}^\kappa(x,y) - \mathrm{k}_M^\kappa(x,y)| \le \overline{K}(F(Q_x) + F(Q_y)) . \tag{29}$$

Proof Denoting $\mathbf{t} = (\mathbf{t}^0, \mathbf{t}^1, \ldots) \coloneqq \mathrm{S}(\kappa_x) \in \mathbf{T}_1(H)$, recall that $\|\mathbf{t}^m\| \leq \|\kappa_x\|_{1-\mathrm{var}}^m / m!$ (Lyons et al., 2007, Prop. 2.2). Let $\pi_M \colon \mathbf{T}(H) \to \prod_{m=0}^{M} H^{\otimes m}$ denote the level-M truncation map. Then

$$\|\pi_M \mathbf{t} - \mathbf{t}\| \leq \Big(\sum_{m>M} \Big(\frac{\|\kappa_x\|_{1-\mathrm{var}}^m}{m!} \Big)^2 \Big)^{1/2} \leq \sum_{m>M} \frac{\|\kappa_x\|_{1-\mathrm{var}}^m}{m!} \leq \frac{e^{\|\kappa_x\|_{1-\mathrm{var}}} \|\kappa_x\|_{1-\mathrm{var}}^{M+1}}{(M+1)!} = Q_x \;.$$

The conclusion now follows from Lemma 45. ∎

References

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

Richard Bass, Ben Hambly, and Terry Lyons. Extending the Wong-Zakai theorem to reversible Markov processes. *J. Eur. Math. Soc. (JEMS)*, 4(3):237–269, 2002.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics.* Springer Science & Business Media, 2011.

Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, USA:, 1994.

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

Richard P. Brent. *Algorithms for minimization without derivatives.* Prentice-Hall, Inc., Englewood Cliffs, N.J., 1973. Prentice-Hall Series in Automatic Computation.

Thomas Cass, Bruce K. Driver, Nengli Lim, and Christian Litterer. On the integration of weakly geometric rough paths. *J. Math. Soc. Japan*, 68(4):1505–1524, 2016.

Antoni B Chan and Nuno Vasconcelos. Probabilistic kernels for the classification of auto-regressive visual processes. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 846–851. IEEE, 2005.

I. Chevyrev, V. Nanda, and H. Oberhauser. Persistence Paths and Signature Features in Topological Data Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(1):192–202, 2020.

Ilya Chevyrev. Random walks and Lévy processes as rough paths. *Probab. Theory Related Fields*, 170(3-4):891–932, 2018.

Ilya Chevyrev and Peter K. Friz. Canonical RDEs and general semimartingales as rough paths. *Ann. Probab.*, 47(1):420–463, 2019.

Ilya Chevyrev and Andrey Kormilitzin. A Primer on the Signature Method in Machine Learning. *ArXiv e-prints*, March 2016.

Ilya Chevyrev and Terry Lyons. Characteristic Functions of Measures on Geometric Rough Paths. *Ann. Probab.*, 44(6):4049–4082, 2016.

Ilya Chevyrev and Marcel Ogrodnik. A support and density theorem for Markovian rough paths. *Electron. J. Probab.*, 23(56):16 pp., 2018.

Laure Coutin and Zhongmin Qian. Stochastic analysis, rough path analysis and fractional Brownian motions. *Probab. Theory Related Fields*, 122(1):108–140, 2002. ISSN 0178-8051.

Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39(1):1–49, 2002.

Marco Cuturi and Arnaud Doucet. Autoregressive kernels for time series. *arXiv preprint arXiv:1101.0673*, 2011.

Thomas Fawcett. *Problems in stochastic analysis: connections between rough paths and non-commutative harmonic analysis*. PhD thesis, University of Oxford, 2003.

Jerome H Friedman and Lawrence C Rafsky. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, pages 697–717, 1979.

Peter Friz and Atul Shekhar. General rough integration, Lévy rough paths and a Lévy–Kintchine type formula. *Ann. Probab.*, 45(4):2707–2765, 2017.

Peter Friz and Nicolas Victoir. On uniformly subelliptic operators and stochastic area. *Probab. Theory Related Fields*, 142(3-4):475–523, 2008.

Peter K. Friz and Martin Hairer. *A course on rough paths*. Universitext. Springer, Cham, 2014. With an introduction to regularity structures.

Peter K. Friz and Nicolas B. Victoir. *Multidimensional stochastic processes as rough paths*, volume 120 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2010.

Peter K. Friz, Benjamin Gess, Archil Gulisashvili, and Sebastian Riedel. The Jain-Monrad criterion for rough paths and applications to random Fourier series and non-Markovian Hörmander theory. *Ann. Probab.*, 44(1):684–738, 2016.

Kenji Fukumizu, Arthur Gretton, Gert R Lanckriet, Bernhard Schölkopf, and Bharath K Sriperumbudur. Kernel choice and classifiability for rkhs embeddings of probability distributions. In *Advances in neural information processing systems*, pages 1750–1758, 2009.

Robin Giles. A generalization of the strict topology. *Transactions of the American Mathematical Society*, 161:467–474, 1971.

Arthur Gretton, Kenji Fukumizu, Zaid Harchaoui, and Bharath K Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in neural information processing systems*, pages 673–681, 2009.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust statistics: the approach based on influence functions*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1986. ISBN 0-471-82921-8.

Peter J. Huber and Elvezio M. Ronchetti. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2009. ISBN 978-0-470-12990-6. doi: 10.1002/9780470434697. URL https://doi.org/10.1002/9780470434697.

Eric Jones, Travis Oliphant, and Pearu Peterson. SciPy: Open source scientific tools for Python, 2001–. URL http://www.scipy.org/.

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

Serge Lang. *Algebra*. Springer,, 2002.

Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

Terry Lyons. Differential equations driven by rough signals. *Rev. Mat. Iberoamericana*, 14(2):215–310, 1998. ISSN 0213-2230.

Terry Lyons and Hao Ni. Expected signature of Brownian motion up to the first exit time from a bounded domain. *Ann. Probab.*, 43(5):2729–2762, 2015.

Terry Lyons and Zhongmin Qian. *System control and rough paths*. Oxford Mathematical Monographs. Oxford University Press, Oxford, 2002. Oxford Science Publications.

Terry Lyons, Michael Caruana, and Thierry Lévy. *Differential equations driven by rough paths*, volume 1908 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007.

Pedro J Moreno, Purdy Ho, and Nuno Vasconcelos. A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications. In *NIPS*, pages 1385–1392, 2003.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

Anastasia Papavasiliou and Christophe Ladroue. Parameter estimation for rough differential equations. *The Annals of Statistics*, 39(4):2047–2073, 2011.

Svetlozar Todorov Rachev. *Probability metrics and the stability of stochastic models*, volume 269. John Wiley & Son Ltd, 1991.

Aaditya Ramdas, Sashank Jakkam Reddi, Barnabás Póczos, Aarti Singh, and Larry A Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *AAAI*, pages 3571–3577, 2015.

Christophe Reutenauer. *Free Lie algebras.* The Clarendon Press Oxford University Press, New York, 1993. ISBN 0-19-853679-8. Oxford Science Publications.

Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT press, 2002.

Carl-Johann Simon-Gabriel and Bernhard Schölkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research*, 19(44):1–29, 2018.

Carl-Johann Simon-Gabriel, Alessandro Barp, and Lester Mackey. Metrizing Weak Convergence with Maximum Mean Discrepancies. *arXiv e-prints*, art. arXiv:2006.09268, June 2020.

Bharath Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839–1893, 2016.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010.

Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12 (Jul):2389–2410, 2011.

Dougal J. Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton. Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy. *arXiv e-prints*, art. arXiv:1611.04488, November 2016.

Csaba Toth and Harald Oberhauser. Bayesian Learning from Sequential Data using Gaussian Processes with Signature Covariances. In *International Conference on Machine Learning*, pages 9548–9560. PMLR, 2020.