Applications of stochastic analysis and algebra to machine learning

Patric Ossian Mauritz Bonnier

St John's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy in Mathematics

2023

Acknowledgements

I would like to express my gratitude to my supervisor Harald Oberhauser for his guidance and support. It has been a privilege to have been your student, and your guidance has been invaluable throughout.

I am very thankful to my examiners, Terry Lyons and Josef Teichmann for agreeing to asses this thesis, and to Zhongmin Qian, Ben Hambly, and Sam Cohen for assessing my transfer and confirmation of status. Your feedback and comments have been very helpful. The support, guidance and help from my collaborators has also been deeply helpful to me. Cristopher Salvi, Imanol Perez Arribas, Patrick Kidger, Csaba Toth, Chong Liu, Zoltán Szabó, and Yannic Vargas, thank you all for your help. I am also thankful to my colleagues and office mates Vlad Margarint, Christina Zou and Alexander Schell for your friendship and support.

I would like to express my gratitude to my family Claës, Yvonne, Isabelle, and Thérèce. Thank you for being there for me and for supporting me this whole time. You are all wonderful, and I am truly blessed to have you in my life. This gratitude gladly extends itself to my friends for supporting my through this journey. You mean more to me than you know, and I am very lucky to have such great people in my life. Lastly, my deepest gratitude goes to my girlfriend Jana and our dog Echo, who have been a constant source of love and support, even if the latter might not fully appreciate this thesis. I love you both.

Abstract

In this thesis we consider the application of tools from stochastic analysis and algebra to statistics and machine learning. Most of these tools are different forms of what has become known as signature methods. The signature has been discovered and rediscovered in a few different areas of mathematics in the last 70 years. In short, it maps a path evolving in a vector space to a group enveloping that same space. The reason it is so useful for statistics is twofold: One, its set of invariances is highly desirable in many applications, and two, its image group is highly structured making it particularly amenable to mathematical study using algebraic tools.

The primary aim here is to study how one may use the signature to express statistical properties of a given path, and how these properties can be applied to machine learning. This aim manifests itself as - among other things:

- a new type of cumulants for signatures that have unique combinatorial properties and can be used to characterise independence of paths,
- cumulants on reproducing kernel Hilbert spaces which are related to the signature cumulants, even though signatures are not used explicitly,
- A generalisation of the signature to other types of feature maps into non-commutative algebras,
- a feature map with an initial topology that captures properties of the filtration of stochastic processes,
- and a family of scoring rules with associated divergences, entropies and mutual informations for paths that respect their group structure.

These are divided into separate, self contained chapters that can be read independently of one another.

Contents

Introduction to this thesis

This thesis is divided into five separate chapters, complete with their own introductions, conclusions and appendices. A cursory glance might give the impression of an eclectic collection of ideas and methods, but they share a strong core in that every chapter is devoted to transforming statistical and analytical problems into algebraic or combinatorial ones and using this perspective to glean new insights and methods. From the start of working on this thesis I knew that I wanted to work more with algebra and combinatorics than is typical for a Dphil in stochastic analysis, and path signatures was an excellent way to do so as it transforms stochastic processes – analytic objects, to random characters on Hopf algebras – algebraic objects. A fresh point of view is often helpful in life, and no less so in Mathematics, and this procedure of transforming statistics to algebra and stochastic analysis proved a fruitful one.

The flavour of each chapter differs, e.g. not every chapter uses path signatures, and some are more analytical than other. However, every chapter in this thesis is singularly devoted to the very same core method of doing Mathematics. Because of this strong common theme, I think that the title of the thesis, while slightly verbose, is very fitting. The chapters are all self-contained and can be read in any order, but the chronological order best represents the evolution of different ideas, from the more combinatorial to the more analytical, and finally the more algebraic, while at the same time roughly representing my own interests and ideas as I grew throughout the Dphil. Next is a brief introduction to the goals and methods of each chapter.

Chapter 1. This chapter aims to extend the classical notion of cumulants to path space in a coordinate free, parametrisation invariant way. Fittingly, this first chapter is also the first one I wrote, and while the algebraic methods employed here are slightly more primitive than in the rest of the thesis I think the combinatorics have a surprising depth. This chapter manages to touch on results far removed from its statistical motivation, and make some connections that I think are still poorly understood in general. Classically, the cumulants of a random variable X are defined as coefficients

of the Taylor expansion of the cumulant generating function

$$t \mapsto \log \mathbb{E} e^{tX}.$$

One way of extending this to path space is to note that the signature $\mathrm{S}(x)$ of a path x behaves similarly to the exponential, which suggests that given a stochastic process, $(X_t)_{t=0}^T$ the series

$$\log \mathbb{E}\, \mathrm{S}(X)$$

would have coefficients with a similar behaviour to the classical cumulants. We show that this is indeed the case, at least in the sense that these "signature cumulants" characterise independence of paths similarly to how cumulants characterise independence of random variables. We do this by studying the combinatorial properties of these coefficients. It is well known that a multivariate cumulant can be expressed as a sum indexed by partitions, and we show that this extends to path space were the sum is taken over what we refer to as "ordered partitions". These ordered partitions have a neat poset interpretations and are largely unexplored as far as we can tell, even if they generalise other common families of partitions. By using these ordered partitions we prove that the signature cumulants characterise independence between paths and use this in some simple experiments to demonstrate their effectiveness.

Chapter 2. This chapter – like the one before it – also aims to take classical cumulants and extend them to new domains. In contrast to Chapter 1, this was the final chapter to be written and proved a nice revisit to the methods employed there but in new settings. This time the extension is to Hilbert spaces, specifically reproducing kernel Hilbert spaces (RKHSs). The reason for this is that some statistics about cumulants on a RKHS can be computed entirely using the reproducing kernel. Using standard tricks from kernel learning one can derive expressions for magnitudes of, and distances between, these "kernelised cumulants". This allows one to write down efficient statistics that characterise independence of probability measures as well as new distances between them. We again use combinatorial expressions for computing these cumulants, but the estimators derived from them are still rather compact. We test these estimators on both real and synthetic data to show how they outperform standard methods in kernel learning.

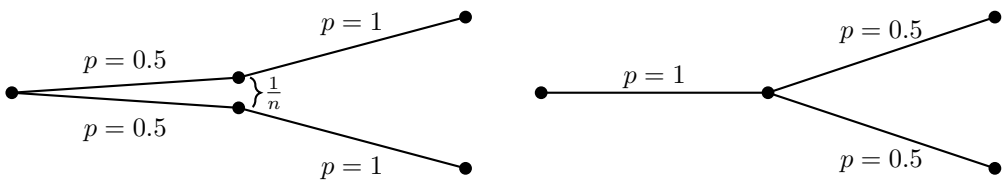
Chapter 3. In this chapter we explore generalisations of the signature mapping of discrete time series by considering homomorphisms from the monoid of paths in a vector space with the concatenation operation into an algebra. That is, maps of the form

$$(v_1, \dots, v_n) \mapsto \varphi(v_1) \star \cdots \star \varphi(v_n)$$

where $v_1, \dots, v_n$ are elements of some vector space V and $\varphi : V \to \mathcal{A}$ takes values in some algebra $(\mathcal{A}, \star)$. We show that under some general assumptions about φ, these maps have similar properties to the signature mapping and is able to capture ordered behaviour when the algebra $\mathcal{A}$ is non-commutative. We also study iterations of such maps in the setting of very simple φ's. These are typically very fast to compute but turn out to be just as expressive as their more complicated counterparts. We use these to build the "neural sequence to tensor" layers and show that they perform well on common datasets. The main text of this chapter was written by Csaba Toth, but it is provided in the thesis for context. I wrote Appendix 3.A, 3.B, and 3.C. The other appendices from the full paper were left out.

Chapter 4. In this chapter, the focus is not on extending some classical statistics to path space, but rather on using signatures to optimise existing methods. We study the *adapted topology* and its sister the *extended weak topology* and their higher order counterparts. We show that by using iterated signatures, one can derive feature maps that induce these topologies (in the sense that the topologies are the initial topologies of the feature maps), as well as metrics for them.

The classical motivation for introducing adapted topologies is the following: In certain applications the filtration associated to a process is important and should be highlighted in the topology one considers. The weak topology does not do this and therefore it is insufficient for these applications. Consider the following toy example taken from Figure 4.1. The process on the left will converge weakly to the one on the



right as $n \to \infty$, but imagine that these are tree models for some financial asset one wishes to invest into, where the nodes represent its price at three different points in time. Using the model on the left one would have complete information about the

future of the asset at the second time point, and using the model on the right one has no more information than at the first time point. This is true regardless of the value of n so in this application these models are fundamentally different.

The idea behind the extended weak topology is to consider the *prediction process* associated to a stochastic process and a filtration $(X, \mathcal{F})$. It is defined as the evolution of the law of the future behaviour conditioned on the present

$$\hat{X}_t := \mathcal{L}(X|\mathcal{F}_t).$$

If one considers the topology induced by weak convergence of these prediction processes, then this yields the extended weak topology, which is strong enough to separate the two toy processes in Figure 4.1.

Our contribution is to show that there is a principled way to compute signatures of prediction processes and that the signature of a higher order prediction process has a simple recursive formulation. These higher order signatures induce the higher order extended weak topologies. For example, the first extended weak topology is induced by the map

$$(X, \mathcal{F}) \mapsto \mathbb{E}\,\mathrm{S}(t \mapsto \mathbb{E}(\mathrm{S}(X)|\mathcal{F}_t))$$

where $\mathrm{S}(\cdot)$ denotes the (augmented) signature. We also provide algorithms for efficient recursive computations for these higher order signatures for the first two non-trivial levels, and show with some simple experiments how they can outperform other methods.

Chapter 5. In this chapter we aim to extend the notion of a *scoring rule* to path space. Scoring rules are a type of discrepancy measure on probability spaces that work differently then other common popular discrepancy measures like the Maximum Mean Discrepancy (MMD). There is a general framework for generating scoring rules, and by using this framework one can build a scoring rule on path space that is parametrisation invariant and respects the group structure of paths (with operations being concatenation and time reversal). One does this by associating each stochastic process X with its so called *Bayes' act* a_X by defining it to be the optimal average inverse of the signature of X.

$$a_X := \min_{a \in \mathcal{H}^\star} \mathbb{E}\|S(X)a\|^2.$$

This in turn leads to the scoring rule between the processes X and a path y being defined as

$$S(X, y) := \|S(y)a_X\|^2.$$

One can leverage this scoring rule to get a divergence, entropy and mutual information for paths and these all have their own respective use cases. For instance, the divergence can be used as a discrepancy measure between probability distributions, and the mutual information can be used as a measure of independence. We show their usefulness with some experiments where we compare them to known baselines.

Chapter 1

Signature cumulants, ordered partitions, and independence of stochastic processes

This chapter is based on [30] which was written with Harald Oberhauser.

The sequence of so-called signature moments describes the laws of many stochastic processes in analogy with how the sequence of moments describes the laws of vector-valued random variables. However, even for vector-valued random variables, the sequence of cumulants is much better suited for many tasks than the sequence of moments. This motivates us to study so-called signature cumulants. To do so, we develop an elementary combinatorial approach and show that in the same way that cumulants relate to the lattice of partitions, signature cumulants relate to the lattice of so-called "ordered partitions". We use this to give a new characterisation of independence of multivariate stochastic processes. Finally, we construct a family of unbiased minimum-variance estimators of signature cumulants and show that even for the simple example of a diffusion with constant drift and volatility, such signature cumulant estimators outperform signature moment estimators.

1.1 Introduction

The sequence of moments $(\mu_X^m)_{m\geq 1}$ of a $\mathbb{R}^d$-valued random variable X,

$$(\mu_X^m)_{m\geq 1} \text{ where } \mu_X^m := \mathbb{E}[X^{\otimes m}] \in (\mathbb{R}^d)^{\otimes m}, \tag{1.1}$$

plays a fundamental role in statistics and probability theory since it captures statistical properties of the law of X in a graded manner. An undesirable property of moments – especially when only a finite number of samples from X is available to estimate them – is that lower order moments can dominate higher order moments. For example, even for $m = 2$ and $d = 1$

$$\mu_X^2 = (\mu_X^1)^2 + \mathrm{Var}(X),$$

and the variance $\mathrm{Var}(X) := \mathbb{E}[(X - \mu_X^1)^2]$ is a much more sensible "order $m = 2$" statistic than μ_2. This method of subtracting lower order moments from higher order moments to obtain better statistics can be continued beyond level $m = 2$; for $m = 3$ one gets the classical notion of "skewness", for $m = 4$ the "tailedness", etc. In general, the resulting sequence is known as the *cumulants*, which differ from the centralised moments in that one allows for subtractions of any lower order moment. Another advantage cumulants have over moments is that many probabilistic properties are often easier expressed in cumulants than in moments.

Example 1.1.1. One good reason to use cumulants instead of moments is that their sample statistics typically have lower variance. To illustrate this on a elementary example assume that $X \sim N(\mu, \sigma^2)$ follows a normal distribution. Given N independent samples $X_1, \ldots, X_N$ of X the minimum variance unbiased estimators for its second moment and cumulant are respectively:

$$\hat{\mu}_X^2 = \frac{1}{N}\sum_{i=1}^N X_i^2, \quad \hat{\kappa}_X^2 = \frac{1}{N-1}\sum_{i=1}^N (X_i - \frac{1}{N}\sum_{j=1}^N X_j)^2$$

By explicit computation in the same fashion as outlined in Appendix 1.C one sees that

$$\mathrm{Var}(\hat{\mu}_X^2) = \mathrm{Var}(\hat{\kappa}_X^2) + \frac{2}{N}\Big(\mu^4 - \mu^2\sigma^2 - \frac{2\sigma^4}{N-1}\Big),$$

It follows that when μ is large compared to σ^2, it is much preferable to use the cumulant estimator rather than the moment estimator.

Example 1.1.2. Another way to think about the difference between estimating moments and cumulants is that by Proposition 1.2.1 one can always write the n:th moment μ_X^n as

$$\mu_X^n = \sum_{i_1+\cdots+i_k=n} \kappa_X^{i_1} \cdots \kappa_X^{i_k},$$

or in other words, the n:th moment is equal to the n:th cumulant plus a polynomial function f_n of lower order terms

$$\mu_X^n = \kappa_X^n + f_n(\mu_X^{n-1}, \ldots, \mu_X^1).$$

This translates immediately to a statement about estimators, since polynomial functions of moment estimators are linear functions on estimators of polynomials

$$\hat{\mu}_X^n = \hat{\kappa}_X^n + \hat{f}_n(\hat{\mu}_X^{n-1}, \ldots, \hat{\mu}_X^1),$$

and since we know by example 1.1.1 that $\hat{\kappa}_X^n$ has a lower variance, Equation 1.1.2 gives weight to the idea that $\hat{\mu}_X^n$ is just $\hat{\kappa}_X^n$ with added noise that depends only on lower order information.

Many real-world observations have an inherent sequential structure and one ultimately deals with path-valued samples (i.e. sample paths of a stochastic process) rather than vector-valued samples. For a path-valued random variable, the sequence of so-called signature moments

$$(\mu_X^m)_{m\geq 0} = \left(\mathbb{E}\left[\int dX^{\otimes m}\right]\right)_{m\geq 0}$$

is a natural replacement for the classical moment sequence (1.1). Indeed, motivated by insights from stochastic analysis [44], these signature moments have recently found applications in statistics and machine learning such as parameter estimation for stochastic differential equations [144] or hypothesis testing for laws of stochastic processes [46]. Motivated by the nice statistical properties of cumulants over moments in the case of vector-valued observations, the aim of this article is to study such properties for signature cumulants and to derive estimators for such signature cumulants. For example, in view of Example 1.1.1 one can ask whether cumulants can yield better estimators for estimating diffusion and drift in an SDE and we invite the reader to have a glance at Example 1.4.3 for an application of our results to a simple SDE with constant drift and volatility. Similarly, signature cumulants between stochastic processes vanish if and only if the stochastic processes are independent which allows for independence testing of time series.

1.1.1 Poset of Partitions, Moments and Cumulants

Before further discussing the path-valued case of signature cumulants, we briefly introduce notation and recall results for classical cumulants below. The following result is classical, see [133] for history and overview. We spell it out in notation unusual in statistics however, as this formulation generalises naturally to our extension. For the rest of the paper we will denote by V a d-dimensional vector space with basis $\{e_1, \dots, e_d\}$. The *tensor algebra* and its topological dual the *extended tensor algebra* of V are defined as

$$\mathrm{T}(V) := \bigoplus_{m\geq 0} V^{\otimes m}, \quad \mathrm{T}((V)) := \prod_{m\geq 0} V^{\otimes m}.$$

The pairing $\langle \cdot, \cdot \rangle$ between $\mathrm{T}(V)$ and $\mathrm{T}((V))$ by $\langle \cdot, \cdot \rangle$ is defined as $\langle s, t \rangle = \sum s_\tau t_\tau$ for $s = \sum s_\tau e_\tau$, $t = \sum t_\tau e_\tau$ where all sums are taken over multiindices $\tau \in \bigcup_m \{1, \dots, d\}^m$ and $e_\tau := e_{i_1} \otimes \cdots \otimes e_{i_m}$.

Theorem 1.1.1 (Classical Cumulants)**.** *Let X be a V-valued random variable such that that the sequence of moments*

$$\mu_X := (\mu_X^m)_{m\geq 0} := \left(\mathbb{E}[X^{\otimes m}]\right)_{m\geq 0} \in \mathrm{T}((V))$$

is well-defined and characterises the law of X. We call[1]

$$\kappa_X := (\kappa_X^m)_{m\geq 0} := \pi_{\mathrm{Sym}}(\log \mathbb{E}[\exp(X)]) \in \mathrm{T}((V)) \tag{1.2}$$

the cumulant of X. For $\tau = (i_1, \dots, i_m) \in \{1, \dots, d\}^m$ it holds that

1. ***(compensated moments)***

$$\langle \kappa_X, e_\tau \rangle = \sum_{\mathbf{a}\in\mathcal{P}(\tau)} (-1)^{|\mathbf{a}|-1}(|\mathbf{a}|-1)! \prod_{i=1}^{k} \langle \mu_X, e_{a_i} \rangle$$

   where the sum ranges over all partitions $\mathbf{a} = \{a_1, \dots, a_k\}$ of τ.

2. ***(moment bijection)*** *There exist a bijection $\kappa_X \mapsto \mu_X$ explicitly given as*

$$\langle \mu_X, e_\tau \rangle = \sum_{\mathbf{a}\in\mathcal{P}(\tau)} \prod_{i=1}^{k} \langle \kappa_X, e_{a_i} \rangle$$

   where again the sum ranges over all partitions of τ.

[1] Using the notation $1 + x_1 + x_2 + \cdots$ for $(x_m)_{m\geq 0}$, the exponential and logarithm are defined as $\exp(x) = \sum_{m\geq 0} \frac{1}{m!} x^{\otimes m}$, $\log(1+x) = \sum_{m\geq 1} \frac{(-1)^{m-1}}{m} x^{\otimes m}$. We denote with $\pi_{\mathrm{Sym}}(v_1 \otimes \cdots \otimes v_m) = \sum_\sigma v_{\sigma(1)} \otimes \cdots \otimes v_{\sigma(m)}$ the projection onto the symmetric tensors, save the $\frac{1}{m!}$ normalising factor.

3. ***(independence)*** *For $I, J \subset \{1, \dots, d\}$ the families $(\langle X, e_i \rangle)_{i \in I}$ and $(\langle X, e_j \rangle)_{j \in J}$ are independent if and only if*

$$\langle \kappa_X, e_{\tau_1} \otimes e_{\tau_2} \rangle = 0 \text{ for any } \tau_1 \in I^\star, \tau_2 \in J^\star$$

where $I^\star := \bigcup_m I^m$ and $J^\star$ is defined similarly.

Item 1 connects cumulants with subtraction of lower moments from higher moments – thus also the term "compensated moments" – with definition (1.2) motivated by the Laplace transform. Item 2 implies that, like the sequence of moments, the sequence of cumulants characterises the law of X. Finally, the third item demonstrates that independence of random variables is equivalent to their cross-cumulants vanishing.

In two seminal papers [171, 172], T. Speed showed that results such as the above can be proven in short and elegant fashion using the lattice of partitions on $\{1, \dots, d\}$. The aim of this paper is to replace probability measures on $\mathbb{R}^d$ by probability measures on spaces of paths and to develop such a combinatorial approach and apply it to derive efficient estimators. As we will see, so-called signature cumulants of laws of stochastic processes are intimately linked to the lattice of "ordered partitions" which has a rich combinatorial structure.

Poset of Ordered Partitions, Signature Moments and Cumulants

Our results apply to any stochastic process for which a "rough path lift" exists, such as continuous semi-martingales. However, even for stochastic processes with smooth sample paths, it is interesting to spell out our "path-space version" of Theorem 1.1.1 which appears below as motivation for the rest of the paper. To speak of moments of measures on paths we first need a notion of monomials of paths. If the path is smooth, this role is taken by iterated integrals. This sequence of tensors is called the *path signature*.

Definition 1.1.1. Define for $x \in \mathrm{BV}(V) := \bigcup_{T>0} \{x \in C([0,T], V) : \|x\|_{1-var} < \infty\}$ its signature $\mathrm{S}(x) = (\mathrm{S}_m(x))_{m \geq 0} = (\int \mathrm{d}x^{\otimes m})_{m \geq 0}$ as

$$\int \mathrm{d}x^{\otimes 0} := 1, \text{ and } \int \mathrm{d}x^{\otimes m} := \int_{0 < t_1 < \cdots < t_m < T} \mathrm{d}x(t_1) \otimes \cdots \otimes \mathrm{d}x(t_m) \quad \text{for } m \geq 1.$$

The study of how such iterated integrals reflect properties of the path x goes back to Chen [41], see also [97]. If one applies this to random paths X, this naturally leads to the notion of signature moments $\mathbb{E}[\mathrm{S}(X)] \in \mathrm{T}((V))$. More recently such moments have found applications in statistics, starting with [144] where a "method

of signature moments" is derived in analogy to the classical method of moments. We state Theorem 1.1.2 to give the reader an idea of the kind of statements we are interested in by using notation defined in the subsequent Section 1.2 and Section 1.3; but even without the concrete definitions of the poset $\mathrm{Orp}(w_1, \dots, w_k)$, the shuffle $\sqcup\!\sqcup$, the unparametrised paths $\mathcal{R}_1(V)$, and the generalised signature moments $\mu_X(\mathbf{a})$ and cumulants $\kappa_X(\mathbf{a})$ at hand, the connection between Theorem 1.1.2 and Theorem 1.1.1 is evident. In fact, as we will in see in Example 1.3.3, Theorem 1.1.1 follows from Theorem 1.1.2 by tensor symmetrisation.

Theorem 1.1.2 (Signature Cumulants)**.** *Let X be a $\mathcal{R}_1(V)$-valued random variable, such that the signature moments*

$$\mu_X := (\mu_X^m)_{m\geq 0} := (\mathbb{E}[\mathrm{S}_m(X)])_{m\geq 0} \in \mathrm{T}(\!(V)\!)$$

are well-defined and characterise the law of X. We call

$$\kappa_X := (\kappa_X^m)_{m\geq 0} := \log \mathbb{E}[\mathrm{S}(X)] \in \mathrm{T}(\!(V)\!)$$

the signature cumulant *of X. For $\tau, \tau_1, \dots, \tau_k \in \bigcup_m \{1, \dots, d\}^m$ it holds that*

1. ***(compensated moments)***

$$\langle \kappa_X, e_{\tau_1} \sqcup\!\sqcup \cdots \sqcup\!\sqcup e_{\tau_k} \rangle = \sum_{\mathbf{a}\in \mathrm{Orp}(\tau_1,\dots,\tau_k)} (-1)^{|\mathbf{a}|-1} \frac{\mathbf{a}!}{|\mathbf{a}|} \mu_X(\mathbf{a}),$$

   where the sum is taken over all ordered partitions of the tuple $(\tau_1, \dots, \tau_k)$,

2. ***(moment bijection)***

$$\langle \mu_X, e_\tau \rangle = \sum_{\mathbf{a}\in \mathrm{Orp}(\tau)} \frac{1}{|\mathbf{a}|!} \kappa_X(\mathbf{a}),$$

   where the sum is taken over all ordered partitions of τ,

3. ***(independence)*** *For $I, J \subset \{1, \dots, d\}$ the families $(\langle X, e_i \rangle)_{i\in I}$ and $(\langle X, e_j \rangle)_{j\in J}$ are independent*[2] *if and only if*

$$\langle \kappa_X, e_{i^\star} \sqcup\!\sqcup e_{j^\star} \rangle = 0 \ \text{ for any } i^\star \in I^\star, j^\star \in J^\star.$$

[2] As usual we call two stochastic processes X and Y independent if their σ-algebras are independent. We emphasize that this is much stronger than instantaneous independence, i.e that X_t and Y_t are independent for every $t \geq 0$.

We emphasise that as for classical cumulants, such results can in principle be derived by direct but lengthy calculations that compare coefficients; or, in contrast, an algebraic approach that relies on the well developed theory of Lie polynomials and Hall bases [152]; or even a Hopf algebraic perspective that has recently been applied to free cumulants [64] as well as Wick polynomials [65]. All of these approaches have their own merits, but as we hope to demonstrate, the combinatorial approach we present here via "ordered partitions" is intuitive, leads to simple proofs, and allows for a self-contained treatment throughout. Möbius inversion and basic properties of U-statistics are the only results we use for which we don't give full proofs. In short, the lattice of ordered partitions is useful for signature cumulants for the same reasons that the lattice of partitions is useful for cumulants.

Remark 1.1.1. There are other notions of non-commutative cumulants such as the Boolean, monotone, and free cumulants that arise in non-commutative probability theory. For instance, the free cumulants are studied using "non-crossing partitions", see [173, 120, 141]. While similar in spirit, our "ordered partitions" have a different combinatorial structure. Yet another related research area is the use of cumulants in Wiener Chaos expansions, see [145] for a survey, but the combinatorics are given by the standard poset of partitions.

Structure

- In Section 1.2 we give some combinatorial background and define the set of ordered partitions. We then go on to show some of their properties in the setting that applies to us.
- In Section 1.3 we give some background on rough paths and study their signature cumulants.
- In Section 1.4 we construct an unbiased minimum variance estimator for signature cumulants and show that already for the simple case of a diffusion with constant drift and volatility, signature cumulants provide more efficient estimators than signature moments.

Appendix 1.A contains technical details of independence of rough paths, Appendix 1.B contains technical information about Tree-like equivalence of paths, and Appendix 1.C contains calculations for Example 1.4.3.

Symbol	Meaning	Page
	Tensors	
V	a d-dimensional vector space	9
$\mathrm{T}(V)$	$\bigoplus V^{\otimes m}$ the free algebra on the vector space V	9
$\mathrm{T}((V))$	$\prod V^{\otimes m}$ the dual of the free algebra $\mathrm{T}(V)$	9
$\langle \cdot, \cdot \rangle$	the pairing between $\mathrm{T}(V)$ and $\mathrm{T}((V))$	9
exp	exponential of a tensor series	9
log	logarithm of a tensor series	9
π_{sym}	the "unormalized" projection to symmetric tensors	9
	Posets and Lattices	
$[n]$	the set $\{1, \dots, n\}$	14
$(P, \leq)$	a poset is a set P with a partial order given by $\leq$	14
$C \subset P$	a totally ordered subset C of P is called a chain	14
$C \subset P$	a subset C of P with no elements comparable is called an antichain	14
m	$m : P \times P \to \mathbb{R}$ the Möbius function	15
	Posets of (ordered) Partitions	
$\mathcal{P}(S)$	poset of partitions of a finite set S	15,16
$\mathcal{P}(d)$	poset of partitions of $\mathcal{P}(S)$ for $S := [d] := \{1, \dots, d\}$	15
$\mathrm{Orp}(P)$	poset of ordered partitions of a finite poset P	17
$\mathrm{P}_{(n_1,\dots,n_k)}$	poset P of disjoint unions of k mutually non-comparable chains C_i of length n_i	17
$\mathbf{a}!$	"factorial" of an ordered partition $\mathbf{a}$ counts the functions that represent the partition $\mathbf{a}$	18
$\mathcal{A}(\mathbf{a})$	antichain ancestry of an ordered partition $\mathbf{a}$	19
$\mathfrak{d}(\mathbf{a})$	"height" of an ordered partition $\mathbf{a}$	19
	Signature Moments and Cumulants	
$\int dx^{\otimes m}$	shorthand for $\int_{0 \leq t_1 \leq \cdots \leq t_m \leq T} dx(t_1) \otimes \cdots \otimes dx(t_m)$	10
$\mathrm{S}(x) = (\mathrm{S}_m(x))_{m \geq 0}$	the signature of continuous bounded variation path x, $\mathrm{S}_m(x) := \int dx^{\otimes m}$	10
$\mathbf{x}$	geometric p-rough path	22
$\mu_{\mathbf{X}} = (\mu^m_{\mathbf{X}})_{m \geq 0}$	signature moments $\mathbb{E}[\mathbf{X}_{0,T}]$ of a random geometric p-rough path $\mathbf{X}$	23
$\kappa_{\mathbf{X}} = (\kappa^m_{\mathbf{X}})_{m \geq 0}$	signature cumulants $\log \mathbb{E}[\mathbf{X}_{0,T}]$ of a random geometric p-rough path $\mathbf{X}$	23
$\mu_{\mathbf{T}}(\mathbf{a})$	generalized moments of a $\mathrm{T}((V))$-valued random variable $\mathbf{T}$	23

$\kappa_{\mathbf{T}}(\mathbf{a})$	generalized cumulants of a $\mathrm{T}(\!(V)\!)$-valued random variable $\mathbf{T}$	23
	Estimators for signature cumulants	
$\hat{k}_n$	Tukey's polykay; min. variance unbiased estimator for classical cumulants	30
$\hat{\kappa}_n$	signature polykay; min. variance unbiased estimator for signature cumulants	32

1.2 The Lattice of (Ordered) Partitions

Given a finite set S, the set of partitions $\mathcal{P}(S)$ of S is a classical combinatorial object. In this section, we show that if S is replaced by a set P that has an additional partial order structure, then a natural generalization of $\mathcal{P}(S)$ are the "ordered partitions" $\mathrm{Orp}(P)$. Analogous to how the poset structure of $\mathcal{P}(S)$ for $S = \{1, \dots, d\}$ plays a role for cumulants of $\mathbb{R}^d$-valued random variables, Theorem 1.1.1, we will see in Section 1.3 that the poset structure of $\mathrm{Orp}(P)$ for $P = \{\tau_1\} \cup \cdots \cup \{\tau_k\}$ with $\tau_1, \dots, \tau_k$ tuples formed from $\{1, \dots, d\}$, plays a similar important role for our signature cumulants, Theorem 1.1.1 and Theorem 1.3.6.

We briefly recall partially ordered sets (henceforth posets), lattices and some important examples such as partitions (classical cumulants), and non-crossing partitions (free cumulants) in Section 1.2.1 and proceed to introduce the lattice of ordered partitions and study its properties in Section 1.2.2.

1.2.1 Posets, Lattices, and Möbius inversion

Throughout this paper we use the notation $[n]$ for the set $\{1, \dots, n\}$ where $n \in \mathbb{N}$. Both $[n]$ and $\mathbb{N}$ are always considered with their normal total order "$\leq$". For a function $f : S \to U$ between sets, we denote by $\ker(f) := \{f^{-1}(u) : u \in U\}$.

A *partially ordered set* (poset) is a pair $(P, \leq)$ where P is a set and "$\leq$" is a partial order on P. It is customary to refer to just P as the poset. A function $f : P_1 \to P_2$ between two posets is said to be *order-preserving* if $f(x) \geq f(y)$ in P_2 whenever $x \geq y$ in P_1, the set of such functions is denoted $\mathrm{Hom}(P_1, P_2)$. We say that a subset $C \subseteq P$ is a *chain* if it is totally ordered and an *antichain* if no element of C is comparable to any other element of C.

An important function on any poset P with finite intervals is the *Möbius function*, defined recursively as:

$$m(\mathbf{a},\mathbf{b}) = \begin{cases} 1, \text{ if } \mathbf{a}=\mathbf{b}, \\ -\sum_{\mathbf{a}\leq\mathbf{c}<\mathbf{b}} m(\mathbf{a},\mathbf{c}) \text{ if } \mathbf{a}<\mathbf{b}, \\ 0, \text{ otherwise.} \end{cases}$$

The most important application is that for finite P, and $f : P \to \mathbb{C}$ and $F : P \to \mathbb{C}$ related by

$$F(\mathbf{a}) = \sum_{\mathbf{b}\leq\mathbf{a}} f(\mathbf{b}),$$

the function f can be recovered from F by "Möbius inversion"

$$f(\mathbf{a}) = \sum_{\mathbf{b}\leq\mathbf{a}} m(\mathbf{b},\mathbf{a})F(\mathbf{b}).$$

If every two elements of $(P,\leq)$ have a unique minimal upper bound and unique maximal lower bound then we say that $(P,\leq)$ is a *lattice*. We refer the reader to [175] for many more results and examples of posets and lattices. Below we recall two lattices that motivate signature cumulants and ordered partitions.

The lattice of partitions. A *partition* $\mathbf{a}$ of a finite set S is a set of disjoint subsets $a_1,\ldots,a_k$ of S such that their union equals S, or equivalently $\mathbf{a} = \ker(f)$ for some $f : S \to \mathbb{N}$. We call the subsets $a_1,\ldots,a_k$ the *blocks* of the partition $\mathbf{a} = \{a_1,\ldots,a_k\}$ and denote by $|\mathbf{a}|$ the number of blocks of $\mathbf{a}$. We denote by $\mathcal{P}(S)$ the *set of partitions of* S and make it into a poset by endowing it with the following partial order: for $\mathbf{a},\mathbf{b} \in \mathcal{P}(S)$

$$\mathbf{a} \leq \mathbf{b}$$

holds if every block of $\mathbf{a}$ is contained in some block of $\mathbf{b}$. In this case, we say that $\mathbf{a}$ is *finer* than $\mathbf{b}$. It is well known that the pair $(\mathcal{P}(S),\leq)$ is a lattice. We will typically denote $\mathcal{P}([d])$ simply by $\mathcal{P}(d)$ and if $\mathbf{a} \in \mathcal{P}(S)$ and $U \subseteq S$ then we use the notation $\mathbf{a}\cap U := \{a\cap U \,|\, a \in \mathbf{a}\} \in \mathcal{P}(U)$. As pointed out in [171], a direct application of Möbius inversion yields a generalisation of the bijection between cumulants and moments.

Proposition 1.2.1 ([171])**.** *For a $\mathbb{R}^d$-valued random variable $X = (X_1,\ldots,X_d)$ define the generalised moments and cumulants of a partition $\mathbf{a} = \{a_1,\ldots,a_k\} \in \mathcal{P}(d)$ as*

$$\mu_X(\mathbf{a}) = \prod_{i=1}^{k} \mathbb{E}(X^{a_i}), \quad \kappa_X(\mathbf{a}) = \prod_{i=1}^{k} \kappa(X_{a_i})$$

where we denote

$$X^a = \prod_{i \in a} X_i, \quad X_a = \times_{i \in a} X_i$$

for a set $a \subseteq [d]$. Then

$$\mu_X(\mathbf{a}) = \sum_{\mathbf{b} \leq \mathbf{a}} \kappa_X(\mathbf{b}), \quad \kappa_X(\mathbf{a}) = \sum_{\mathbf{b} \leq \mathbf{a}} \mu_X(\mathbf{b}) m(\mathbf{b}, \mathbf{a}).$$

In the special case where $\mathbf{a}$ is the partition with only one block, Proposition 1.2.1 reduces to the well known formula

$$\kappa(X_1, \ldots, X_d) = \sum_{\mathbf{a} \in \mathcal{P}(d)} (-1)^{|\mathbf{a}|-1} (|\mathbf{a}| - 1)! \mu_X(\mathbf{a})$$

which is Item 1 of Theorem 1.1.1. We point the reader to [171, 172] for many more results that can be derived with this approach.

The lattice of non-crossing partitions. If a set S is totally ordered, then a partition $\mathbf{a} \in \mathcal{P}(S)$ is said to be *crossing* if there exists $x_1 < x_2 < x_3 < x_4$ in S such that x_1, x_3 and x_2, x_3 belong to two distinct blocks of $\mathbf{a}$. If one requires that partitions be compatible with the order on S in sense of not being crossing, then one arrives at the set of *non-crossing partitions* of S, typically denoted by $NC(S)$. When endowed with the partial order inherited from $\mathcal{P}(S)$, $NC(S)$ becomes a lattice. This lattice naturally appears in free probability where, if one requires that Proposition 1.2.1 holds, one arrives at the *free cumulants*.

1.2.2 The lattice of ordered partitions

We now introduce a lattice with a richer structure by introducing *ordered partitions*. In spite of the fact that they are defined in a natural way, the literature on them is limited. To the best of the authors knowledge they first appear in the paper [179] by T. Sturm, and a similar notion appears in the PhD-thesis of R. Stanley [176]. More recently they have found use in combinatorial algebraic topology, [106]. In this section we derive some basic results that we will need in the sequel; in Section 1.3, we see that they are naturally linked to signature cumulants.

Recall that set of partitions of a finite set S can be defined as follows:

Definition 1.2.1. Let S be a finite set. We call

$$\mathcal{P}(S) := \{\ker(f) \mid f : S \to \mathbb{N}\}$$

the set of partions of S.

In view of this, the following definition is natural

Definition 1.2.2. Let $(P, \leq)$ be a finite poset. We call

$$\mathrm{Orp}(P) := \{\ker(f) \mid f \in \mathrm{Hom}(P, \mathbb{N})\}$$

the set of ordered partions of P.

Equipped with the partial order of refinement both $\mathcal{P}(P)$, and its subset $\mathrm{Orp}(P)$, form a lattice [179, Theorem 21], and the inclusion map $\iota : \mathrm{Orp}(P) \to \mathcal{P}(P)$ is order-preserving. Like in the case of non-crossing partitions, $\mathrm{Orp}(P)$ is not a sublattice of $\mathcal{P}(P)$ in general.

Example 1.2.1. See Figure 1.1 for some simple examples; in addition

1. Take $P = \{x_1, x_2, y_1, y_2\}$ with $x_1 \leq x_2$ and $y_1 \leq y_2$ and no other relations. Then the partition $\{x_1, y_1\}, \{x_2, y_2\}$ is ordered since the map $f : P \to \{1, 2\}, f(x_1) = f(y_1) = 1, f(x_2) = f(y_2) = 2$ is order-preserving. The partition $\{x_1, y_2\}, \{x_2, y_1\}$ is not ordered however since if f is order-preserving such that $f(x_1) = f(y_2)$ and $f(y_1) = f(x_2)$ then we would need $f(x_2) \geq f(x_1) = f(y_2) \geq f(y_1) = f(x_2)$, hence $\ker(f) \neq \{\{x_1, y_2\}, \{x_2, y_1\}\}$.

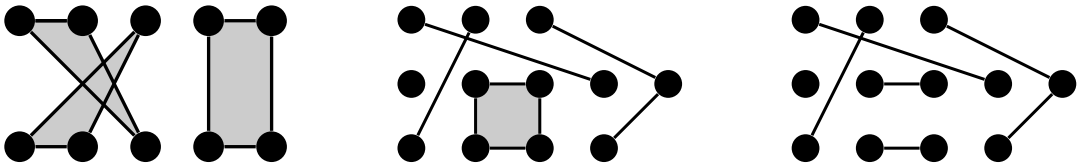
2. If P is an antichain, then $\mathrm{Orp}(P) = \mathcal{P}(P)$.

3. If P is totally ordered, then $\mathrm{Orp}(P)$ is the set of partitions of P with convex blocks, i.e if x and y belong to the same block then so do all $x \leq z \leq y$.

Remark 1.2.1. Given a partition $\mathbf{a} \in \mathcal{P}(P)$, we say that a sequence $x_0, \ldots x_n$ is an $\mathbf{a}$-*cycle* if $x_0 = x_n$ and for every $i \in \{0, \ldots, n-1\}$ either $x_i < x_{i+1}$ or x_i, x_{i+1} belong to the same block of $\mathbf{a}$. If we say that a partition $\mathbf{a} \in \mathcal{P}(P)$ is n-*ordered* if $x_0, \ldots, x_n$ all belong to the same block of $\mathbf{a}$ for any $\mathbf{a}$-cycle $x_0, \ldots, x_n$. Then we get the characterisation that $\mathbf{a}$ is ordered if it is n-ordered for every $n \geq 1$ ([180, Lemma 2]) and non-crossing if it's 4-ordered.

In this paper we are mainly concerned with P a disjoint union of mutually non-comparable chains as defined below:

Definition 1.2.3. We denote by $\mathrm{P}_{(n_1,\ldots,n_k)}$ the poset $P = C_1 \cup \cdots \cup C_k$ where each C_i is a chain of length n_i and $x \in C_i, y \in C_j$ are comparable if and only if $i = j$.



(a) Not an ordered partition (b) Not an ordered partition (c) An ordered partition

Figure 1.1: Using the notation from Definition 1.2.3, one element of $\mathrm{P}_{(5,5)}$ and two elements of $\mathrm{P}_{(3,5,4)}$ all drawn with the order going left to right. One can check that (1.1a),(1.1b) are not ordered partitions while (1.1c) is.

For a partition $\mathbf{a} \in \mathcal{P}(S)$ the number of blocks is related to the number of functions that represent it via the factorial,

$$|\mathbf{a}|! = \#\{f : S \to [|\mathbf{a}|] \mid \ker(f) = \mathbf{a}\}.$$

This is no longer true for ordered partitions and motivates the following notation

Definition 1.2.4. For $\mathbf{a} \in \mathrm{Orp}(P)$, we define its *factorial* $\mathbf{a}!$ as

$$\mathbf{a}! := \#\{f \in \mathrm{Hom}(P, [|\mathbf{a}|]) \Big| \ker(f) = \mathbf{a}\}.$$

For this next example, recall that if $C \subseteq P$ and $\mathbf{a} \in \mathcal{P}(P)$ we use the notation

$$\mathbf{a} \cap C := \{a \cap C \mid a \in \mathbf{a}\} \in \mathcal{P}(C)$$

Example 1.2.2.

1. If $P = C_1 \cup \cdots \cup C_k = \mathrm{P}_{(|C_1|,\ldots,|C_k|)}$, and $\mathbf{a} = (\mathbf{a} \cap C_1) \cup \cdots \cup (\mathbf{a} \cap C_k)$ is a disjoint union of its intersection with each chain, then $\mathbf{a}!$ is easily seen to be:
$$\mathbf{a}! = \frac{|\mathbf{a}|!}{\prod_{i=1}^{k} |\mathbf{a} \cap C_i|!}.$$
2. If P is an antichain, then $\mathbf{a}! = |\mathbf{a}|!$ for any $\mathbf{a} \in \mathrm{Orp}(P)$.

Remark 1.2.2. Counting the number of order-preserving maps from a poset P to a chain $[n]$ is a classic topic in combinatorics and computer science, where one is typically concerned with bijective maps, so-called *linear extensions*. An important quantity in order theory is the order polynomial $\Omega(n, P)$ and strict order polynomial $\Omega^\circ(n, P)$ of P, see [33]. They will not be used here but we remark that for a finite poset P and $n \leq \#P$

$$\sum_{\mathbf{a} \in \mathrm{Orp}(P), |\mathbf{a}| \leq n} \mathbf{a}! = \Omega(n, P), \qquad \sum_{\mathbf{a} \in \mathrm{Orp}(P), |\mathbf{a}| = n} \mathbf{a}! = \Omega^\circ(n, P),$$

which follows directly from the definition.

Another important quantity that appears in the context of signature cumulants is the *antichain ancestry* of $\mathbf{a} \in \mathrm{Orp}(P)$.

Definition 1.2.5. For $\mathbf{a} \in \mathrm{Orp}(P)$ we define the set

$$\mathcal{A}(\mathbf{a}) = \{\mathbf{b} \geq \mathbf{a} : \mathbf{b} \cap C = \mathbf{a} \cap C \text{ for any chain } C \subseteq P\}.$$

We further define $\mathfrak{d}(\mathbf{a})$ as:

$$\mathfrak{d}(\mathbf{a}) = \sum_{\mathbf{b} \in \mathcal{A}(\mathbf{a})} (-1)^{|\mathbf{b}|-1} \frac{\mathbf{b}!}{|\mathbf{b}|}.$$

As it turns out, there is a correspondence between $\mathcal{A}(\mathbf{a})$ and $\{f \in \mathrm{Hom}(P, [|\mathbf{a}|]) \Big| \ker(f) = \mathbf{a}\}$, and $\mathfrak{d}(\mathbf{a}) = 0$ whenever $\mathbf{a}$ is minimal in its antichain ancestry. This is spelled out below in the case where P is the disjoint union of exactly two chains.

Remark 1.2.3. If P is an antichain then this is easy to see since

$$\mathfrak{d}(\mathbf{a}) = \sum_{\mathbf{a} \leq \mathbf{b}} (-1)^{|\mathbf{b}|-1}(|\mathbf{b}|-1)! = \sum_{\mathbf{a} \leq \mathbf{b} \leq \mathbb{1}} m(\mathbf{b}, \mathbb{1}) = \begin{cases} 1 \text{ if } \mathbf{a} = \mathbb{1} \\ 0 \text{ otherwise,} \end{cases}$$

where $\mathbb{1}$ is the partition with one block.

Proposition 1.2.2. *If* $P = C_1 \cup C_2 = \mathrm{P}_{(|C_1|,|C_2|)}$ *and* $\mathbf{a} \in \mathrm{Orp}(P)$*, then*

$$\mathbf{a}! = \#\mathcal{A}(\mathbf{a}).$$

Moreover, if $\mathbf{a} = (\mathbf{a} \cap C_1) \cup (\mathbf{a} \cap C_2)$*, then* $\mathfrak{d}(\mathbf{a}) = 0$*.*

Proof. Assume without loss of generality that $\mathbf{a} \cap C_i = C_i$. Define the set

$$\mathcal{L}(\mathbf{a}) = \{f \in \mathrm{Hom}(P, [|\mathbf{a}|]) \Big| \ker(f) = \mathbf{a}\}.$$

If $\{x, y\} \subseteq C_1 \times C_2$ and $\{u, v\} \subseteq C_1 \times C_2$ are two blocks of $\mathbf{a}$ then either $x \leq u, y \leq v$ or $u \leq x, v \leq y$. Hence if $f, g \in \mathcal{L}(\mathbf{a})$ and $\{x, y\} \subseteq C_1 \times C_2$ is a block of $\mathbf{a}$, then $f(x) = f(y) = g(x) = g(y)$. So if we let $\mathbf{a}_0^\star, \ldots, \mathbf{a}_k^\star$ be the blocks of $\mathbf{a}$ with exactly two elements and $\mathbf{a}_i$ be the (possibly empty) collection of elements x_i such that $\mathbf{a}_{i-1}^\star \cap C_1 < x_i < \mathbf{a}_i^\star \cap C_1$ if $x_i \in \mathbf{a}_i \cap C_1$ and $\mathbf{a}_{i-1}^\star \cap C_2 < x_i < \mathbf{a}_i^\star \cap C_2$ if $x_i \in \mathbf{a}_i \cap C_2$, then we may identify:

$$\mathcal{L}(\mathbf{a}) = \mathcal{L}(\mathbf{a}_1) \times \cdots \times \mathcal{L}(\mathbf{a}_k).$$

Additionally, if $\{x, y\} \subseteq C_1 \times C_2$ is a block of $\mathbf{a}$ and $\mathbf{b} \in \mathcal{A}(\mathbf{a})$ has a block $\{u, v\} \subseteq C_1 \times C_2$, then either:

1. $u = x, v = y$
2. $u < x, v < y$
3. $u > x, v > y$

In view of this we may make the similar identification:

$$\mathcal{A}(\mathbf{a}) = \mathcal{A}(\mathbf{a}_1) \times \cdots \times \mathcal{A}(\mathbf{a}_k).$$

To show the first claim we may therefore assume that $\mathbf{a} = C_1 \cup C_2$ so that the blocks of $\mathbf{a}$ are singletons, but this is now clear since in this case, if $\mathbf{b} \in \mathcal{A}(\mathbf{a})$, then $\mathbf{b}$ is the aggregation of some number of blocks of C_1 with the same number of blocks in C_2. Since $\mathbf{b}$ is ordered, there is exactly one unique way of aggregating them when the blocks are chosen. So there's exactly $\binom{|C_1|}{m}\binom{|C_2|}{m}$ number of ordered partitions $\mathbf{b} \in \mathcal{A}(\mathbf{a})$ that aggregates m blocks of $\mathbf{a}$, hence

$$\#\mathcal{A}(\mathbf{a}) = \sum_{m=0}^{|C_1|\wedge|C_2|} \binom{|C_1|}{m}\binom{|C_2|}{m} = \frac{|C|!}{|C_1|!|C_2|!} = \mathbf{a}!$$

by Remark 1. This shows the first claim. To see the second claim, define the auxiliary sets

$$\mathcal{A}_n(\mathbf{a}) = \{\mathbf{b} \in \mathcal{A}(\mathbf{a}) : |\mathbf{b}| = |\mathbf{a}| - n\}$$

and let $z_n(\mathbf{a}) = \#\mathcal{A}_n(\mathbf{a})$. We know that:

$$\mathbf{a}! = \sum_{n\geq 0} z_n(\mathbf{a}).$$

Noting the identity

$$\sum_{\mathbf{b}\in\mathcal{A}_i(\mathbf{a})} z_j(\mathbf{b}) = \binom{i+j}{i} z_{i+j}(\mathbf{a}),$$

we can write

$$\begin{aligned}
\mathfrak{d}(\mathbf{a}) &= \sum_{\mathbf{b}\geq\mathbf{a}} (-1)^{|\mathbf{b}|-1} \frac{\mathbf{b}!}{|\mathbf{b}|} \\
&= \sum_{n\geq 0} \frac{1}{|\mathbf{a}|-n} (-1)^{|\mathbf{a}|-n-1} \sum_{\mathbf{b}\in\mathcal{A}_n(\mathbf{a})} \sum_{k\geq 0} z_k(\mathbf{b}) \\
&= (-1)^{|\mathbf{a}|-1} \sum_{n\geq 0} \frac{1}{|\mathbf{a}|-n} (-1)^n \sum_{k\geq 0} \binom{n+k}{k} z_{n+k}(\mathbf{a}) \\
&= (-1)^{|\mathbf{a}|-1} \sum_{m\geq 0} z_m(\mathbf{a}) \sum_{n+k=m} \frac{1}{|\mathbf{a}|-n} (-1)^n \binom{m}{k} \\
&= (-1)^{|\mathbf{a}|-1} \sum_{m\geq 0} (-1)^m \frac{z_m(\mathbf{a})}{(|\mathbf{a}|-m)\binom{|\mathbf{a}|}{m}}.
\end{aligned}$$

Now fix $\mathbf{a} = (\mathbf{a} \cap C_1) \cup (\mathbf{a} \cap C_2)$ and let $|\mathbf{a} \cap C_1| = m_1$ and $|\mathbf{a} \cap C_2| = m_2$. By the same reasoning as before we have

$$z_m(\mathbf{a}) = \binom{m_1}{m}\binom{m_2}{m},$$

and we get

$$\begin{aligned}
\mathfrak{d}(\mathbf{a}) &= (-1)^{|\mathbf{a}|-1} \sum_{m=0}^{m_1 \wedge m_2} (-1)^m \frac{\binom{m_1}{m}\binom{m_2}{m}}{(m_1+m_2-m)\binom{m_1+m_2}{m}} \\
&= (-1)^{|\mathbf{a}|-1} \frac{1}{m_1+m_2} \sum_{m=0}^{m_1 \wedge m_2} (-1)^m \frac{\binom{m_1}{m}\binom{m_2}{m}}{\binom{m_1+m_2-1}{m}} \\
&= (-1)^{|\mathbf{a}|-1} \frac{1}{m_1+m_2} {}_2F_1(-m_1, -m_2; -m_1-m_2+1, 1) = 0
\end{aligned}$$

where ${}_2F_1$ is the Gaussian hypergeometric function. □

1.3 The Signature Cumulant and its properties

The lattice of ordered partitions from Section 1.2 provides us with the combinatorial language to prove properties about signature cumulants in analogy with classical cumulants. Section 1.3.1 recalls geometric rough paths and in Section 1.3.2 we introduce generalised signature moments and cumulants. Finally, in Section 1.3.3 we prove Theorem 1.3.6 that is our "path-space" version of the classical cumulant results, Theorem 1.1.1.

1.3.1 Geometric rough paths

Given a basis $\{e_1, \dots, e_d\}$ for a vector space V, the set $\{e_\tau := e_{i_1} \otimes \cdots \otimes e_{i_m} : \tau = (i_1, \dots, i_m) \in [d]^m\}$ is a basis for $V^{\otimes m}$. Throughout, we denote with $[d]^\star = \bigcup_{m \geq 0} [d]^m$ the set of tuples of arbitrary length.

Definition 1.3.1. The shuffle product $\sqcup\!\sqcup : \mathrm{T}(V) \times \mathrm{T}(V) \to \mathrm{T}(V)$ is defined by linear extension of the map

$$e_{\tau_1} \sqcup\!\sqcup e_{\tau_2} = \sum_{\tau \in Sh(\tau_1, \tau_2)} e_\tau, \quad \tau_1, \tau_2 \in [d]^\star$$

where the sum is taken over all shuffles[3] of τ_1 and τ_2.

[3] A shuffle of tuples $\tau_1 = (i_1, \dots, i_m) \in [d]^m$ and $\tau_2 = (j_1, \dots, j_n) \in [d]^n$ is given by mapping $(\sigma_1, \dots, \sigma_{m+n}) := (i_1, \dots, i_m, j_1, \dots, j_n)$ to $(\sigma_{\pi(1)}, \dots, \sigma_{\pi(m+n)})$ where π is a permutation on $[m+n]$ such that $\pi(1) < \cdots \pi(m)$ and $\pi(m+1) < \cdots < \pi(m+n)$.

Remark 1.3.1. If $\tau_1 = (i_1, \dots, i_m), \tau_2 = (i_{m+1}, \dots, i_{m+n})$ then their shuffles can equivalently be defined as:

$$\mathrm{Sh}(\tau_1, \tau_2) = \{f(\tau_1, \tau_2) = (i_{f^{-1}(1)}, \dots, i_{f^{-1}(m+n)}) \mid \text{bijections } f \in \mathrm{Hom}(\mathrm{P}_{(m,n)}, \mathrm{P}_{(m+n)})\}.$$

Definition 1.3.2. Given $p \geq 1$, a *weakly geometric p-rough path* in V is a map $\mathbf{x} : [0,T]^2 \to \mathrm{T}((V))$ such that for all $0 \leq s < t < u \leq T$

1. $\langle \mathbf{x}_{s,t}, f \sqcup\!\sqcup g \rangle = \langle \mathbf{x}_{s,t}, f \rangle \langle \mathbf{x}_{s,t}, g \rangle$ for all $f, g \in \mathrm{T}(V)$,
2. $\mathbf{x}_{s,t} \otimes \mathbf{x}_{t,u} = \mathbf{x}_{s,u}$,
3. $\|\mathbf{x}\|_{p-var} < \infty$.

where

$$\|\mathbf{x}\|_{p-var} = \max_{1 \leq m \leq p} \sup_{D} \Big(\sum_{t_i \in D} |\mathbf{x}^m_{t_i, t_{i+1}}|^{p/m} \Big)^{1/p},$$

and the sum is taken over all dissections[4] of $[0,T]$. The space of weakly geometric p-rough paths in V is denoted by $\mathcal{G}^w_p(V)$, and $\mathcal{G}^w_p(V)$-valued random variables will be denoted by $\mathbf{X}$.

Random rough paths. Our main interest here is that many stochastic processes X in the state space V can be naturally lifted to random (weakly) geometric rough paths $\mathbf{X}$.

Example 1.3.1.

1. If the sample paths of X have finite p-variation for some $p < 2$ then the lift may be constructed using Young integration. For instance, any piecewise C^1 path has finite 1-variation.
2. If X is a continuous semi-martingale then the lift may be constructed using Stratonovich integration.
3. If X is a Gaussian process with i.i.d. components and a covariance function of finite p-variation for some $p < 4$, then there is a canonical lift of X, see [76].

[4]By a *dissection* of $[0,T]$ we mean a finite collection $t_1 < \cdots < t_{|D|} \in [0,T]$

1.3.2 Signature moments and cumulants.

In view of Theorem 1.1.1, the following definition is not surprising

Definition 1.3.3. We call

$$\mu_{\mathbf{X}} := (\mu_{\mathbf{X}}^m)_{m\geq 0} := \mathbb{E}[\mathbf{X}_{0,T}] \in \mathrm{T}(\!(V)\!)$$

the *signature moments* of $\mathbf{X}$ and we call

$$\kappa_{\mathbf{X}} := (\kappa_{\mathbf{X}}^m)_{m\geq 0} := \log \mathbb{E}[\mathbf{X}_{0,T}] \in \mathrm{T}(\!(V)\!)$$

the *signature cumulant* of $\mathbf{X}$ whenever these quantities are well-defined.

It is well known that the signature moments characterise the law of a signature if they decay fast enough, see [44]. Throughout the remainder of Section 1.3, $\mathbf{X}$ denotes a geometric rough path for which the signature moments are well-defined.

Generalised signature cumulants As in the classical case, a generalisation of moments and cumulants will be useful; see e.g. [171, 172] and Proposition 1.2.1. Applied to tuples that are single indices the definition below recovers the classical generalized moments and cumulants.

Definition 1.3.4. Let $\tau_1, \dots, \tau_k \in [d]^\star$. We denote with $\mathcal{P}(\tau_1, \dots, \tau_k)$ the partitions of the set $\{\tau_1\} \cup \cdots \cup \{\tau_k\}$ and $\mathrm{Orp}(\tau_1, \dots, \tau_k)$ the ordered partitions of the same set endowed with the poset structure of $\mathrm{P}_{(|\tau_1|,\dots,|\tau_k|)}$. For $\mathbf{a} = \{a_1, \dots, a_n\} \in \mathcal{P}(\tau_1, \dots, \tau_k)$, we define $a_i^j := \tau_j \cap a_i$.

Definition 1.3.5. Given a random variable $\mathbf{T}$ with values in $\mathrm{T}(\!(V)\!)$ and $\mathbf{a} \in \mathcal{P}(\tau_1, \dots, \tau_k)$, we call

$$\mu_{\mathbf{T}}(\mathbf{a}) = \prod_{i=1}^{|\mathbf{a}|} \mathbb{E}\Big[\langle \mathbf{T}, e_{a_i^1}\rangle \cdots \langle \mathbf{T}, e_{a_i^k}\rangle\Big]$$

the generalised $\mathbf{a}$-moments of $\mathbf{T}$ and

$$\kappa_{\mathbf{T}}(\mathbf{a}) = \prod_{i=1}^{|\mathbf{a}|} \kappa\Big(\langle \mathbf{T}, e_{a_i^1}\rangle, \dots, \langle \mathbf{T}, e_{a_i^k}\rangle\Big)$$

the generalised $\mathbf{a}$-cumulants of $\mathbf{T}$. Here, by convention, $a_i^1, \dots, a_i^k$ only runs over the non-empty a_i^j's.

Example 1.3.2. If $\mathbf{a} = 1235|67|4 \in \mathcal{P}(1234, 567)$ then

$$\mu_{\mathbf{T}}(\mathbf{a}) = \mathbb{E}\Big(\langle \mathbf{T}, e_{123}\rangle\langle \mathbf{T}, e_5\rangle\Big)\mathbb{E}\Big(\langle \mathbf{T}, e_{67}\rangle\Big)\mathbb{E}\Big(\langle \mathbf{T}, e_4\rangle\Big),$$
$$\kappa_{\mathbf{T}}(\mathbf{a}) = \kappa\Big(\langle \mathbf{T}, e_{123}\rangle, \langle \mathbf{T}, e_5\rangle\Big)\kappa\Big(\langle \mathbf{T}, e_{67}\rangle\Big)\kappa\Big(\langle \mathbf{T}, e_4\rangle\Big).$$

Lemma 1.3.1. *For a random variable* $\mathbf{T}$ *with values in* $\mathrm{T}((V))$ *and* $\mathbf{a} \in \mathcal{P}(\tau_1, \ldots, \tau_k)$*, it holds that*

$$\mu_{\mathbf{T}}(\mathbf{a}) = \sum_{\mathbf{b} \leq \mathbf{a}, \mathbf{a} \in \mathcal{A}(\mathbf{b})} \kappa_{\mathbf{T}}(\mathbf{b}).$$

Proof. To see this, fix $\mathbf{a} \in \mathcal{P}(\tau_1, \ldots, \tau_k)$ and let P be the poset $a^1 \cup \cdots \cup a^k$ with an antichain ordering. Let $\Phi : \{\tau_1\} \cup \cdots \cup \{\tau_k\} \to P$ be the map such that $\Phi(x) = a^i_j$ if $x \in a^i_j$. Given a random variable $\mathbf{T}$, define the random variable $\mathbf{T}'$ taking values in $\mathbb{R}^P$ by

$$\langle \mathbf{T}', e_i\rangle := \langle \mathbf{T}, e_{\Phi^{-1}(i)}\rangle, \quad i \in P$$

then we can then write

$$\mu_{\mathbf{T}}(\mathbf{a}) = \mu_{\mathbf{T}'}(\Phi(\mathbf{a})),$$

where Φ acts on $\mathbf{a}$ by

$$\Phi(\mathbf{a}) = \{\Phi(a) : a \in \mathbf{a}\} \in \mathcal{P}(P).$$

Note that $\mu_{\mathbf{T}'}(\Phi(\mathbf{a}))$ is the moment as defined in Equation 1.2.1 since P is an antichain. By Proposition 1.2.1 we get

$$\mu_{\mathbf{T}'}(\Phi(\mathbf{a})) = \sum_{\mathbf{b}' \leq \Phi(\mathbf{a})} \kappa_{\mathbf{T}'}(\mathbf{b}')$$

which, by definition of $\mathbf{T}'$ yields

$$\mu_{\mathbf{T}}(\mathbf{a}) = \sum_{\mathbf{b} = \Phi^{-1}(\mathbf{b}'), \mathbf{b}' \leq \Phi(\mathbf{a})} \kappa_{\mathbf{T}}(\mathbf{b}).$$

So it is enough to note that $\mathbf{b} = \Phi^{-1}(\mathbf{b}')$ with $\mathbf{b}' \leq \Phi(\mathbf{a})$ if and only if $\mathbf{b} \leq \mathbf{a}$ and $\mathbf{a} \in \mathcal{A}(\mathbf{b})$. □

1.3.3 Cumulants, moments, and independence

We show Theorem 1.1.2, restated for weakly geometric rough paths at the end of this section.

Lemma 1.3.2. *For any $p \geq 1$ and $\mathcal{G}_p^w(V)$-valued random variable $\mathbf{X}$ and any $\tau \in [d]^\star$*

$$\langle \kappa_{\mathbf{X}}, e_\tau \rangle = \sum_{\mathbf{a} \in \mathrm{Orp}(\tau)} \frac{(-1)^{|\mathbf{a}|-1}}{|\mathbf{a}|} \mu_{\mathbf{X}}(\mathbf{a}),$$
$$\langle \mu_{\mathbf{X}}, e_\tau \rangle = \sum_{\mathbf{a} \in \mathrm{Orp}(\tau)} \frac{1}{|\mathbf{a}|!} \kappa_{\mathbf{X}}(\mathbf{a}).$$

Proof. Let $\overline{\mu_{\mathbf{X}}} = \mu_{\mathbf{X}} - 1$, i.e

$$\langle \overline{\mu_{\mathbf{X}}}, e_\tau \rangle = \begin{cases} \langle \mu_{\mathbf{X}}, e_\tau \rangle \text{ if } \tau \neq \varnothing, \\ 0 \text{ otherwise.} \end{cases}$$

We may write

$$\log \mu_{\mathbf{X}} = \log\left(1 + \overline{\mu_{\mathbf{X}}}\right) = \sum_{n \geq 1} \frac{(-1)^{n-1}}{n} \left(\overline{\mu_{\mathbf{X}}}\right)^{\otimes n} = \sum_{n \geq 1} \frac{(-1)^{n-1}}{n} \Big(\sum_{m \geq 1} \mu_{\mathbf{X}}^m \Big)^{\otimes n}$$
$$\mu_{\mathbf{X}} = \exp \kappa_X = 1 + \sum_{n \geq 1} \frac{1}{n!} \kappa_X^{\otimes n} = 1 + \sum_{n \geq 1} \frac{1}{n!} \Big(\sum_{m \geq 1} \kappa_X^m \Big)^{\otimes n}.$$

By identifying coordinates we see that:

$$\langle \kappa_{\mathbf{X}}, e_\tau \rangle = \sum_{n \geq 1} \sum_{\tau_1 \cdots \tau_n = \tau} \frac{(-1)^{n-1}}{n} \prod_i^n \langle \mu_{\mathbf{X}}, e_{\tau_i} \rangle,$$
$$\langle \mu_{\mathbf{X}}, e_\tau \rangle = \sum_{n \geq 1} \sum_{\tau_1 \cdots \tau_n = \tau} \frac{1}{n!} \prod_i^n \langle \kappa_{\mathbf{X}}, e_{\tau_i} \rangle$$

where the sum is over all de-concatenations of τ into non-empty sub-words $\tau_1, \ldots, \tau_k$. The assertion now follows from the identification made in Item 3 of example 1.2.1. □

Proposition 1.3.3. *For any $p \geq 1$ and $\mathcal{G}_p^w(V)$-valued random variable $\mathbf{X}$ and any tuple $\tau = (\tau_1, \ldots, \tau_k)$, $\tau_1, \ldots, \tau_k \in [d]^\star$*[5]

$$\langle \kappa_{\mathbf{X}}, e_{\tau_1} ш \cdots ш e_{\tau_k} \rangle = \sum_{\mathbf{a} \in \mathrm{Orp}(\tau)} (-1)^{|\mathbf{a}|-1} \frac{\mathbf{a}!}{|\mathbf{a}|} \mu_{\mathbf{X}}(\mathbf{a})$$

[5]To avoid having to deal with partitions on multisets we will assume throughout that all indices are distinct. This is treated similarly in [171, 172], that is, all indices are treated as distinct even if the same index appears multiple times.

Proof. Let $N = |\tau_1| + \cdots + |\tau_k|$ and write $|\tau| = (|\tau_1|, \ldots, |\tau_k|)$ for brevity. Define the set

$$\mathrm{Sh}(\tau) = \{f(\tau) \mid \text{ bijective } f \in \mathrm{Hom}(\mathrm{P}_{|\tau|}, [N])\}.$$

By Lemma 1.3.2 and Remark 1.3.1 we may write:

$$\langle \kappa_X, e_{\tau_1} \sqcup\!\sqcup \cdots \sqcup\!\sqcup e_{\tau_k} \rangle = \sum_{w \in \mathrm{Sh}(\tau)} \sum_{\mathbf{a} \in \mathrm{Orp}(w)} \frac{(-1)^{|\mathbf{a}|-1}}{|\mathbf{a}|} \mu_{\mathbf{X}}(\mathbf{a}) = \sum_{\mathbf{a} \in \mathrm{Orp}(N)} \frac{(-1)^{|\mathbf{a}|-1}}{|\mathbf{a}|} \sum_{w \in \mathrm{Sh}(\tau)} \mu_{\mathbf{X}}(w|\mathbf{a})$$

where we write $w|\mathbf{a}$ for the partition of τ with blocks $(w|\mathbf{a})_i = \bigcup_{j \in a_i} w_j$. So it is enough to show that for any $q \in \mathbb{N}$

$$\sum_{\substack{\mathbf{a} \in \mathrm{Orp}(N) \\ |\mathbf{a}| = q}} \sum_{w \in \mathrm{Sh}(\tau)} \mu_{\mathbf{X}}(w|\mathbf{a}) = \sum_{\substack{\mathbf{b} \in \mathrm{Orp}(\tau) \\ |\mathbf{b}| = q}} \mathbf{b}! \mu_{\mathbf{X}}(\mathbf{b}).$$

Fix some $\mathbf{a} = \ker(f_{\mathbf{a}}) = \{a_1, \ldots, a_q\} \in \mathrm{Orp}(N)$ and some $w \in \mathrm{Sh}(\tau)$. Since $[N]$ is totally ordered $f_{\mathbf{a}} : [N] \to [q]$ is uniquely determined by $\mathbf{a}$. Since $w \in \mathrm{Sh}(\tau)$ there exists some linear extension $f_w : \mathrm{P}_{|\tau|} \to [N]$ so that $w = f_w(\tau)$. Define a map

$$\Phi : \mathrm{Orp}(N) \times \mathrm{Sh}(\tau) \to \mathrm{Hom}(\mathrm{P}_{|\tau|}, [q]), \quad (\mathbf{a}, w) \mapsto f_{\mathbf{a}} \circ f_w.$$

In addition, for any $\mathbf{a} \in \mathrm{Orp}(N)$, $f_{\mathbf{b}} \in \mathrm{Hom}(\mathrm{P}_{|\tau|}, [q])$, define the set:

$$\mathcal{D}(\mathbf{a}, f_{\mathbf{b}}) := \left\{ w \in \mathrm{Sh}(\tau) \mid f_{\mathbf{b}} = f_{\mathbf{a}} \circ f_w \right\}$$

and let $H(f_{\mathbf{b}})$ be the unique element of $\mathrm{Orp}(N)$ such that $\# f_{\mathbf{a}}^{-1}(i) = \# f_{\mathbf{b}}^{-1}(i)$ for every $i \in [q]$. Note the following:

1. Φ is surjective since if $f_{\mathbf{b}} \in \mathrm{Hom}(\mathrm{P}_{|\tau|}, [q])$, then by fixing some linear extensions on every block of $\mathbf{b}$ we may factor $f_{\mathbf{b}} = \tilde{f}_{\mathbf{b}} \circ h$ where $h : \mathrm{P}_{|\tau|} \to [N]$ is a linear extension of $\mathrm{P}_{|\tau|}$. Then $\Phi(\ker\left(\tilde{f}_{\mathbf{b}}\right), h(\tau)) = f_{\mathbf{b}}$.

2. The kernel of Φ decomposes as follows:

$$\Phi^{-1}(f_{\mathbf{b}}) = \left\{ (\mathbf{a}, w) \in \mathrm{Orp}(N) \times \mathrm{Sh}(\tau) \,\middle|\, \mathbf{a} = H(f_{\mathbf{b}}), w \in \mathcal{D}(\mathbf{a}, f_{\mathbf{b}}) \right\}.$$

3. For any $f_{\mathbf{b}} \in \mathrm{Hom}(\mathrm{P}_{|\tau|}, [q])$ such that $\ker(f_{\mathbf{b}}) = \mathbf{b}$ and $\mathbf{a} = H(f_{\mathbf{b}})$, item 1 of Definition 1.3.2 implies that

$$\sum_{w \in \mathcal{D}(\mathbf{a}, f_{\mathbf{b}})} \mu_{\mathbf{X}}(w|\mathbf{a}) = \mu_{\mathbf{X}}(\mathbf{b}).$$

In view of this we can write

$$\begin{aligned}
&\sum_{\substack{\mathbf{a}\in\mathrm{Orp}(N)\\|\mathbf{a}|=q}}\sum_{w\in\mathrm{Sh}(\tau)}\mu_{\mathbf{X}}(w|\mathbf{a})\\
&=\sum_{f_{\mathbf{b}}\in\mathrm{Hom}(\mathrm{P}_{|\tau|},[q])}\ \sum_{(\mathbf{a},w)\in\Phi^{-1}(f_{\mathbf{b}})}\mu_{\mathbf{X}}(w|\mathbf{a})\\
&=\sum_{f_{\mathbf{b}}\in\mathrm{Hom}(\mathrm{P}_{|\tau|},[q])}\ \sum_{w\in\mathcal{D}(H(f_{\mathbf{b}}),\mathbf{b})}\mu_{\mathbf{X}}(w|H(f_{\mathbf{b}}))\\
&=\sum_{f_{\mathbf{b}}\in\mathrm{Hom}(\mathrm{P}_{|\tau|},[q])}\mu_{\mathbf{X}}(\mathbf{b})\\
&=\sum_{\substack{\mathbf{b}\in\mathrm{Orp}(\tau)\\|\mathbf{b}|=q}}\mathbf{b}!\mu_{\mathbf{X}}(\mathbf{b}).
\end{aligned}$$

□

Corollary 1.3.4. *Under the same assumptions as Proposition 1.3.3:*

$$\langle\kappa_{\mathbf{X}},e_{\tau_1}\shuffle\cdots\shuffle e_{\tau_k}\rangle=\sum_{\mathbf{a}\in\mathrm{Orp}(\tau)}\mathfrak{d}(\mathbf{a})\kappa_{\mathbf{X}}(\mathbf{a}).$$

Proof. Using Remark 1.3.1 we may write

$$\begin{aligned}
\langle\kappa_X,e_{\tau_1}\shuffle\cdots\shuffle e_{\tau_k}\rangle&=\sum_{\mathbf{a}\in\mathrm{Orp}(\tau)}(-1)^{|\mathbf{a}|-1}\frac{\mathbf{a}!}{|\mathbf{a}|}\mu_{\mathbf{X}}(\mathbf{a})\\
&=\sum_{\mathbf{a}\in\mathrm{Orp}(\tau)}(-1)^{|\mathbf{a}|-1}\frac{\mathbf{a}!}{|\mathbf{a}|}\sum_{\mathbf{b}\leq\mathbf{a},\mathbf{a}\in\mathcal{A}(\mathbf{b})}\kappa_{\mathbf{X}}(\mathbf{b})\\
&=\sum_{\mathbf{b}\in\mathrm{Orp}(\tau)}\kappa_{\mathbf{X}}(\mathbf{b})\sum_{\mathbf{a}\in\mathcal{A}(\mathbf{b})}(-1)^{|\mathbf{a}|-1}\frac{\mathbf{a}!}{|\mathbf{a}|}\\
&=\sum_{\mathbf{b}\in\mathrm{Orp}(\tau)}\mathfrak{d}(\mathbf{b})\kappa_{\mathbf{X}}(\mathbf{b}).
\end{aligned}$$

□

Example 1.3.3. If $\tau=(i_1,\dots,i_m)\in[d]^\star$, then by Remark 1.2.3 we can conclude that:

$$\langle\pi_{\mathrm{sym}}(\kappa_{\mathbf{X}}),e_\tau\rangle=\langle\kappa_{\mathbf{X}},e_{i_1}\shuffle\cdots\shuffle e_{i_m}\rangle=\kappa(\langle\mathbf{X}_{0,T},e_{i_1}\rangle,\dots,\langle\mathbf{X}_{0,T},e_{i_m}\rangle),$$

and one retrieves the classical cumulant on the increment $\mathbf{X}_{0,T}$.

Proposition 1.3.5. *Let $I,J\subseteq[d]$ be two sets of coordinates and let $\mathbf{X}$ take values in $\mathcal{G}_p^w(V)$ for some $p\geq1$. Then*

1. $\mathbb{E}\Big[\langle \mathbf{X}_{0,T}, e_{\tau_1}\rangle\langle \mathbf{X}_{0,T}, e_{\tau_2}\rangle\Big] = \langle \mu_{\mathbf{X}}, e_{\tau_1}\rangle\langle \mu_{\mathbf{X}}, e_{\tau_2}\rangle, \quad \forall\, \tau_1 \in I^\star, \tau_2 \in J^\star$
   if and only if

2. $\langle \kappa_{\mathbf{X}}, e_{\tau_1} \sqcup\!\sqcup e_{\tau_2}\rangle = 0, \quad \forall\, \tau_1 \in I^\star, \tau_2 \in J^\star$ *non-empty,*

where $I^\star := \{(i_1, \dots, i_m) \,|\, i_1, \dots, i_m \in I, m \geq 1\}$.

Proof. For any fixed $\tau_1 \in I^\star, \tau_2 \in J^\star$ we say that $\mathbf{a} \in \mathrm{Orp}(\tau_1, \tau_2)$ is non-degenerate if $\mathbf{a} \neq (\mathbf{a} \cap \tau_1) \cup (\mathbf{a} \cap \tau_2)$. By Corollary 1.3.4 and Proposition 1.2.2 we may write

$$\langle \kappa_{\mathbf{X}}, e_{\tau_1} \sqcup\!\sqcup e_{\tau_2}\rangle = \kappa\Big(\langle \mathbf{X}_{0,T}, e_{\tau_1}\rangle, \langle \mathbf{X}_{0,T}, e_{\tau_2}\rangle\Big) + \sum_{\substack{\mathbf{a}\in \mathrm{Orp}(\tau_1,\tau_2), |\mathbf{a}|\geq 2 \\ \mathbf{a} \text{ non-degenerate}}} \mathfrak{d}(\mathbf{a})\kappa_{\mathbf{X}}(\mathbf{a}).$$

Noting that 1 is equivalent to $\kappa\Big(\langle \mathbf{X}_{0,T}, e_{\tau_1}\rangle, \langle \mathbf{X}_{0,T}, e_{\tau_2}\rangle\Big) = 0$ for every $\tau_1 \in I^\star, \tau_2 \in J^\star$ non-empty, the assertion follows by induction on the length of τ_1, τ_2. □

Remark 1.3.2. We would like to thank the referee for pointing out the following algebraic proof of Proposition 1.3.5 which relies on the bialgebra structure of $\mathrm{T}((V))$ instead of the combinatorial results established here:

Write $\mathbb{R}^I, \mathbb{R}^J$ for the vector spaces spanned by $\{e_i \,|\, i \in I\}$ and $\{e_j \,|\, j \in J\}$ respectively and denote by $\pi_I : \mathbb{R}^d \to \mathbb{R}^I, \pi_J : \mathbb{R}^d \to \mathbb{R}^J$ the associated projection maps. These naturally extend to projection maps $\pi_I^\star : \mathrm{T}((\mathbb{R}^d)) \to \mathrm{T}((\mathbb{R}^I)), \pi_J^\star : \mathrm{T}((\mathbb{R}^d)) \to \mathrm{T}((\mathbb{R}^J))$ which are both graded algebra morphisms. Denoting by Δ the map defined by algebraic extension of $\Delta v = 1 \otimes v + v \otimes 1$, it is well known that Δ is a graded algebra morphism and is dual to $\sqcup\!\sqcup$ in the sense that

$$\langle x, f \sqcup\!\sqcup g\rangle = \langle \Delta x, f \otimes g\rangle$$

for $x \in \mathrm{T}((V)), f, g \in \mathrm{T}(V^\star)$, see e.g [152, Proposition 1.9]. Hence Item 2 is equivalent to

$$\langle \Delta\, \kappa_{\mathbf{X}}, e_{\tau_1} \otimes e_{\tau_2}\rangle = \langle \kappa_{\mathbf{X}} \otimes 1 + 1 \otimes \kappa_{\mathbf{X}}, e_{\tau_1} \otimes e_{\tau_2}\rangle$$

for any $\tau_1 \in I^\star, \tau_2 \in J^\star$; The above can be restated as

$$(\pi_I^\star \otimes \pi_J^\star)\Delta\, \kappa_{\mathbf{X}} = (\pi_I^\star \otimes \pi_J^\star)(\kappa_{\mathbf{X}} \otimes 1 + 1 \otimes \kappa_{\mathbf{X}}).$$

Since $\pi_I^\star \otimes \pi_J^\star : \mathrm{T}((\mathbb{R}^d)) \otimes \mathrm{T}((\mathbb{R}^d)) \to \mathrm{T}((\mathbb{R}^I)) \otimes \mathrm{T}((\mathbb{R}^J))$ and $\Delta : \mathrm{T}((\mathbb{R}^d)) \to \mathrm{T}((\mathbb{R}^d)) \otimes \mathrm{T}((\mathbb{R}^d))$ are graded algebra morphisms they factor through the exponential map, which we may apply to both sides of Equation 1.3.2 to get

$$(\pi_I^\star \otimes \pi_J^\star)\Delta e^{\kappa_{\mathbf{X}}} = (\pi_I^\star \otimes \pi_J^\star) e^{\kappa_{\mathbf{X}} \otimes 1 + 1 \otimes \kappa_{\mathbf{X}}} = (\pi_I^\star \otimes \pi_J^\star)(e^{\kappa_{\mathbf{X}}} \otimes e^{\kappa_{\mathbf{X}}})$$

since $e^{\kappa_{\mathbf{X}}} = \mu_{\mathbf{X}}$ by definition, this is equivalent to

$$(\pi_I^\star \otimes \pi_J^\star)\Delta\, \mu_{\mathbf{X}} = (\pi_I^\star \otimes \pi_J^\star)(\mu_{\mathbf{X}} \otimes \mu_{\mathbf{X}})$$

which is equivalent to Item 1.

To summarise this section, we have now proven Theorem 1.1.2. To state it appropriately we first need to define the space of *unparametrised* weakly geometric rough paths.

Definition 1.3.6. The space of *unparametrised* weakly geometric p-rough paths is defined as the quotient space

$$\mathcal{R}_p^w(V) = \mathcal{G}_p^w(V)/\sim$$

where $\sim$ denotes so called *tree-like equivalence* of rough paths and is essentially factoring out different ways of parametrising it. See Appendix 1.B for details.

Remark 1.3.3. It is well known in the literature that $\mathbf{X} \mapsto \mathbf{X}_{0,T}$ is a bijection for $\mathbf{X} \in \mathcal{R}_p^w(V)$ [23, Theorem 1.1].

Putting everything together and using the notation $\mathbf{X}|_I = (\langle \mathbf{X}, e_\tau\rangle)_{\tau \in I^\star}$ for a set of coordinates $I \subseteq [d]$ shows the following result

Theorem 1.3.6. *Let $p \geq 1$ and $\mathbf{X}$ be a $\mathcal{R}_p^w(V)$-valued random variable such that $\mathbb{E}[|\langle \mathbf{X}_{0,T}, f\rangle|] < \infty$ for every $f \in \mathrm{T}(V)$. Then for $\tau, \tau_1, \dots, \tau_k \in [d]^\star$ it holds that*

1. ***(compensated moments)***

$$\langle \kappa_{\mathbf{X}}, e_{\tau_1} \sqcup\!\sqcup \cdots \sqcup\!\sqcup e_{\tau_k}\rangle = \sum_{\mathbf{a} \in \mathrm{Orp}(\tau_1,\dots,\tau_k)} (-1)^{|\mathbf{a}|-1} \frac{\mathbf{a}!}{|\mathbf{a}|} \mu_{\mathbf{X}}(\mathbf{a}),$$

2. ***(moment bijection)***

$$\langle \mu_{\mathbf{X}}, e_\tau\rangle = \sum_{\mathbf{a} \in \mathrm{Orp}(\tau)} \frac{1}{|\mathbf{a}|!} \kappa_{\mathbf{X}}(\mathbf{a}),$$

3. ***(independence)*** *If $I, J \subseteq [d]$ and the joint law $(\mathbf{X}|_I, \mathbf{X}|_J)$ is characterised by*

$$\Big\{\mathbb{E}[\langle \mathbf{X}_{0,T}, e_{\tau_1}\rangle \langle \mathbf{X}_{0,T}, e_{\tau_2}\rangle] \,\Big|\, \tau_1 \in I^\star, \tau_2 \in J^\star\Big\},$$

then $\mathbf{X}|_I$ and $\mathbf{X}|_J$ are independent if and only if

$$\langle \kappa_{\mathbf{X}}, e_{\tau_1} \sqcup\!\sqcup e_{\tau_2}\rangle = 0 \ \text{ for any } \tau_1 \in I^\star, \tau_2 \in J^\star.$$

A sufficient condition for Item 3 is that the individual signature moments decay sufficiently fast, see Appendix 1.A. This can be verified for many models; an alternative is to use "normalised signatures", [46], for which the normalised signature moments always characterise the law which is useful for "black box" approaches as used in machine learning.

1.4 Estimating Signature cumulants

To apply the previous results in practice, one has to take into account that only a finite number of sample paths are available and needs to derive efficient estimators for signature cumulants. For classical cumulants, Tukey introduce so-called *Polykays* in [192]. We recall these in Section 1.4.1 and present them in a formulation and show that they are indeed the minimum variance unbiased estimators in Lemma 1.4.1 (this is probably well-known but we could not find a reference). In Section 1.4.2 we then introduce such polykays for signature cumulants and show their optimality and other properties as estimators in Proposition 1.4.2. Finally, we apply these results to give an application to drift and volatility estimation for SDEs in Example 1.4.3.

1.4.1 Tukey's Polykays

Given a partition $\mathbf{a} = \{a_1, \dots, a_k\}$ of $[d]$, the *symmetric means* are polynomials in $n \times d$ variables $\left(x_i^1, \dots, x_i^d\right)_{i=1}^n$ defined as follows:

$$\hat{\mu}_n(\mathbf{a}) = \frac{1}{(n)_k} \sum_{1 \le i_1, i_2, \dots, i_k \le n}^{\neq} x_{i_1}^{a_1} \cdots x_{i_k}^{a_k}$$

where $(n)_k = n(n-1)\cdots(n-k)$, $x_i^S = \prod_{j \in S} x_i^j$ and the notation $\sum^{\neq}$ means that the sum is taken over all non-equal indices. The *Polykays* – originally introduced by Tukey in [192] – are defined as

$$\hat{k}_n(\mathbf{a}) = \sum_{\mathbf{b} \le \mathbf{a}} m(\mathbf{b}, \mathbf{a}) \hat{\mu}_n(\mathbf{b})$$

where m is the Möbius function of the partition lattice. See e.g [172, 133] for more on polykays.

Example 1.4.1.

$$\hat{\mu}_n(12) = \frac{1}{n} \sum_{i=1}^n x_i^1 x_i^2, \quad \hat{\mu}_n(1|2) = \frac{1}{n(n-1)} \sum_{1 \le j \ne i \le n} x_i^1 x_j^2.$$

For $\mathbf{a} \in \mathcal{P}(m)$ and i.i.d random variables $(X_i)_{i=1}^n = \left(X_i^1, \dots, X_i^m\right)_{i=1}^n$ the symmetric means and polykays are unbiased estimators of moments and cumulants respectively, i.e

$$\mathbb{E}\hat{\mu}_n(\mathbf{a}) = \mu_X(\mathbf{a}), \quad \mathbb{E}\hat{k}_n(\mathbf{a}) = \kappa_X(\mathbf{a}).$$

Lemma 1.4.1. *For any finite disjoint sets S_1, S_2 and $\mathbf{a}_1 \in \mathcal{P}(S_1), \mathbf{a}_2 \in \mathcal{P}(S_2)$:*

$$\mathrm{Cov}(\hat{k}_n(\mathbf{a}_1), \hat{k}_n(\mathbf{a}_2)) = \frac{1}{n} \sum_{\mathbf{c} \in V(\mathbf{a}_1, \mathbf{a}_2)} \sum_{\substack{\mathbf{b} \leq \mathbf{c} \\ \mathbf{b} \nmid \mathbf{a}_1, \mathbf{b} \nmid \mathbf{a}_2}} \kappa_X(\mathbf{b}) + \mathcal{O}(n^{-2})$$

where

$$V(\mathbf{a}_1, \mathbf{a}_2) = \{\mathbf{c} \in \mathcal{P}(S_1 \cup S_2) : \mathbf{c} \cap S_1 = \mathbf{a}_1, \mathbf{c} \cap S_2 = \mathbf{a}_2, |\mathbf{c}| = |\mathbf{a}_1| + |\mathbf{a}_2| - 1\},$$

and the notation $\mathbf{b} \nmid \mathbf{a}$ means that no block of $\mathbf{b}$ is a proper subset of a block of $\mathbf{a}$.

Proof. It follows from [172, Proposition 3.1] that for any $\mathbf{a}_1 \in \mathcal{P}(S_1), \mathbf{a}_2 \in \mathcal{P}(S_2)$

$$\mathbb{E}(\hat{k}_n(\mathbf{a}_1)\hat{k}_n(\mathbf{a}_2)) = \sum_{\mathbf{b} \in \mathcal{P}(S_1 \cup S_2)} c(\mathbf{b}) \kappa_X(\mathbf{b}),$$
$$c(\mathbf{b}) = \sum_{\mathbf{c} \geq \mathbf{b}} \frac{(n)_{\mathbf{c}}}{(n)_{\mathbf{c}_1}(n)_{\mathbf{c}_2}} m(\mathbf{c}_1, \mathbf{a}_1) m(\mathbf{c}_2, \mathbf{a}_2)$$

where $(n)_{\mathbf{c}} = n(n-1)\cdots(n-|\mathbf{c}|)$ and $\mathbf{c}_i = \mathbf{c} \cap S_i$. By [172, Proposition 3.2] we know that $c(\mathbf{b}) = 0$ if any block of $\mathbf{b}$ is a proper subset of a block of $\mathbf{a}_1$ or $\mathbf{a}_2$, hence the only $\mathcal{O}(1)$ term is $\kappa_X(\mathbf{a}_1)\kappa_X(\mathbf{a}_2)$.

Unless there exists some $\mathbf{c} \geq \mathbf{b}$ such that $|\mathbf{c}| = |\mathbf{c}_1| + |\mathbf{c}_2| - 1$, then $c(\mathbf{b}) \in \mathcal{O}(n^{-2})$ and $\frac{(n)_{\mathbf{c}}}{(n)_{\mathbf{c}_1}(n)_{\mathbf{c}_2}} = \frac{1}{n} + \mathcal{O}(n^{-2})$ if this is the case. Additionally, since no block of $\mathbf{b}$ is a proper subset of a block of $\mathbf{a}_1, \mathbf{a}_2$ this must also be true for such a $\mathbf{c} \geq \mathbf{b}$. Combined with the constraint $|\mathbf{c}| = |\mathbf{c}_1| + |\mathbf{c}_2| - 1$ this implies that $\mathbf{c}_1 = \mathbf{a}_1, \mathbf{c}_2 = \mathbf{a}_2$, so $\mathbf{c} \in V(\mathbf{a}_1, \mathbf{a}_2)$. In view of this we may write:

$$\mathrm{Cov}(\hat{k}_n(\mathbf{a}_1), \hat{k}_n(\mathbf{a}_2)) = \frac{1}{n} \sum_{\substack{\mathbf{b} \in \mathcal{P}(S_1 \cup S_2) \\ \mathbf{b} \nmid \mathbf{a}_1, \mathbf{b} \nmid \mathbf{a}_2}} \kappa_X(\mathbf{b}) \sum_{\mathbf{c} \geq \mathbf{b}, \mathbf{c} \in V(\mathbf{a}_1, \mathbf{a}_2)} m(\mathbf{c}_1, \mathbf{a}_1) m(\mathbf{c}_2, \mathbf{a}_2) + \mathcal{O}(n^{-2})$$
$$= \frac{1}{n} \sum_{\mathbf{c} \in V(\mathbf{a}_1, \mathbf{a}_2)} \sum_{\substack{\mathbf{b} \leq \mathbf{c} \\ \mathbf{b} \nmid \mathbf{a}_1, \mathbf{b} \nmid \mathbf{a}_2}} \kappa_X(\mathbf{b}) + \mathcal{O}(n^{-2}).$$

□

Remark 1.4.1. If one is interested in computing the magnitude of the error in the $\mathcal{O}(n^{-2})$ term then one would use the formula from [172],

$$\mathrm{Cov}(\hat{k}_n(\mathbf{a}_1)\hat{k}_n(\mathbf{a}_2)) = \sum_{\mathbf{b}} c(\mathbf{b}) \kappa_X(\mathbf{b}), \quad c(\mathbf{b}) = \sum_{\mathbf{c} \geq \mathbf{b}} \frac{(n)_{\mathbf{c}}}{(n)_{\mathbf{c}_1}(n)_{\mathbf{c}_2}} m(\mathbf{c}_1, \mathbf{a}_1) m(\mathbf{c}_2, \mathbf{a}_2)$$

and sum over all $\mathbf{b} \in \mathcal{P}(S_1 \cup S_2)$ such that $\frac{(n)_{\mathbf{c}}}{(n)_{\mathbf{c}_1}(n)_{\mathbf{c}_2}} \in \mathcal{O}(n^{-2})$ for all $\mathbf{c} \geq \mathbf{b}$.

Example 1.4.2. If $\mathbf{a} \in \mathcal{P}(S)$ is a partition with blocks of size at most 2, then Lemma 1.4.1 shows that:

$$\mathrm{Var}(\hat{k}_n(\mathbf{a})) = \frac{1}{n} \sum_{a_i, a_j \in \mathbf{a}} \Big(\kappa_X(a_i a_j) + \kappa_X(a_i^1 a_j^1)\kappa_X(a_i^2 a_j^2) + \kappa_X(a_i^1 a_j^2)\kappa_X(a_i^2 a_j^1) \Big) \prod_{k \neq i,j} \kappa_X(a_k)^2 + \mathcal{O}(n^{-2}).$$

Definition 1.4.1. If $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim \mathbf{X}$ is an i.i.d. sequence of random variables in $\mathcal{G}_p^w(V)$ and $\tau = (\tau_1, \ldots, \tau_n)$ then we define

$$\hat{\kappa}_n(\tau) = \sum_{\mathbf{a} \in \mathrm{Orp}(\tau)} (-1)^{|\mathbf{a}|-1} \frac{\mathbf{a}!}{|\mathbf{a}|} \hat{\mu}_n(\mathbf{a}) = \sum_{\mathbf{a} \in \mathrm{Orp}(\tau)} \mathfrak{d}(\mathbf{a}) \hat{k}_n(\mathbf{a}).$$

1.4.2 Signature polykays

The following Proposition describes properties of $\hat{\kappa}_n(\tau)$. Recall that for a function f of k variables and i.i.d. samples $X_1, \ldots, X_n \sim X$, the *U-statistic* of $f(X)$ is defined as

$$U(f)_n = \mathbb{E}_{\sigma \in \mathrm{Sym}(n)}[f(X_{\sigma(1)}, \ldots, X_{\sigma(k)})]$$

where the expectation is taken over the uniform measure of permutations $\sigma : [n] \to [n]$, see [195].

Proposition 1.4.2. *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim \mathbf{X}$ be an i.i.d sequence such that $\mathbb{E}|\langle \mathbf{X}, e_\tau \rangle| < \infty$ for every $\tau \in [d]^\star$, then for any $\tau = (\tau_1, \ldots, \tau_k)$:*

1. $\mathbb{E}\hat{\kappa}_n(\tau) = \langle \kappa_{\mathbf{X}}, e_{\tau_1} \sqcup\!\sqcup \cdots \sqcup\!\sqcup e_{\tau_k} \rangle$,
2. $\hat{\kappa}_n(\tau)$ *is minimum variance in the family of unbiased polynomial estimators of* $\langle \kappa_{\mathbf{X}}, e_{\tau_1} \sqcup\!\sqcup \cdots \sqcup\!\sqcup e_{\tau_k} \rangle$,
3. $\hat{\kappa}_n(\tau) \to \langle \kappa_{\mathbf{X}}, e_{\tau_1} \sqcup\!\sqcup \cdots \sqcup\!\sqcup e_{\tau_k} \rangle$ *a.s. and in L^p for every $p \in [1, \infty)$ as $n \to \infty$,*
4. *For any finite collection S of index tuples, the family $\Big(\hat{\kappa}_n(\tau)\Big)_{\tau \in S}$ is asymptotically normal with asymptotic covariance $\mathcal{V}(\tau_1, \tau_2) = \sum_{\substack{\mathbf{a}_1 \in \mathrm{Orp}(\tau_1) \\ \mathbf{a}_2 \in \mathrm{Orp}(\tau_2)}} \mathfrak{d}(\mathbf{a}_1)\mathfrak{d}(\mathbf{a}_2)\mathcal{V}(\mathbf{a}_1, \mathbf{a}_2)$,*

where $\mathcal{V}(\mathbf{a}_1, \mathbf{a}_2) = \sum_{\mathbf{c} \in V(\mathbf{a}_1, \mathbf{a}_2)} \sum_{\substack{\mathbf{b} \leq \mathbf{c} \\ \mathbf{b} \nmid \mathbf{a}_1, \mathbf{b} \nmid \mathbf{a}_2}} \kappa_{\mathbf{X}}(\mathbf{b})$.

Proof. U-statistics are minimum variance in the family of unbiased polynomial estimators, converging a.s. and in L^p, as well as asymptotically normal [195, Section 12]. Moreover, if U, V are U-statistics for $f(X), g(X)$, then $U + V$ and (U, V) are U-statistics for $(f + g)(X)$ and $(f, g)(X)$ respectively. Since $\hat{k}_n(\mathbf{a})$ is a U-statistic

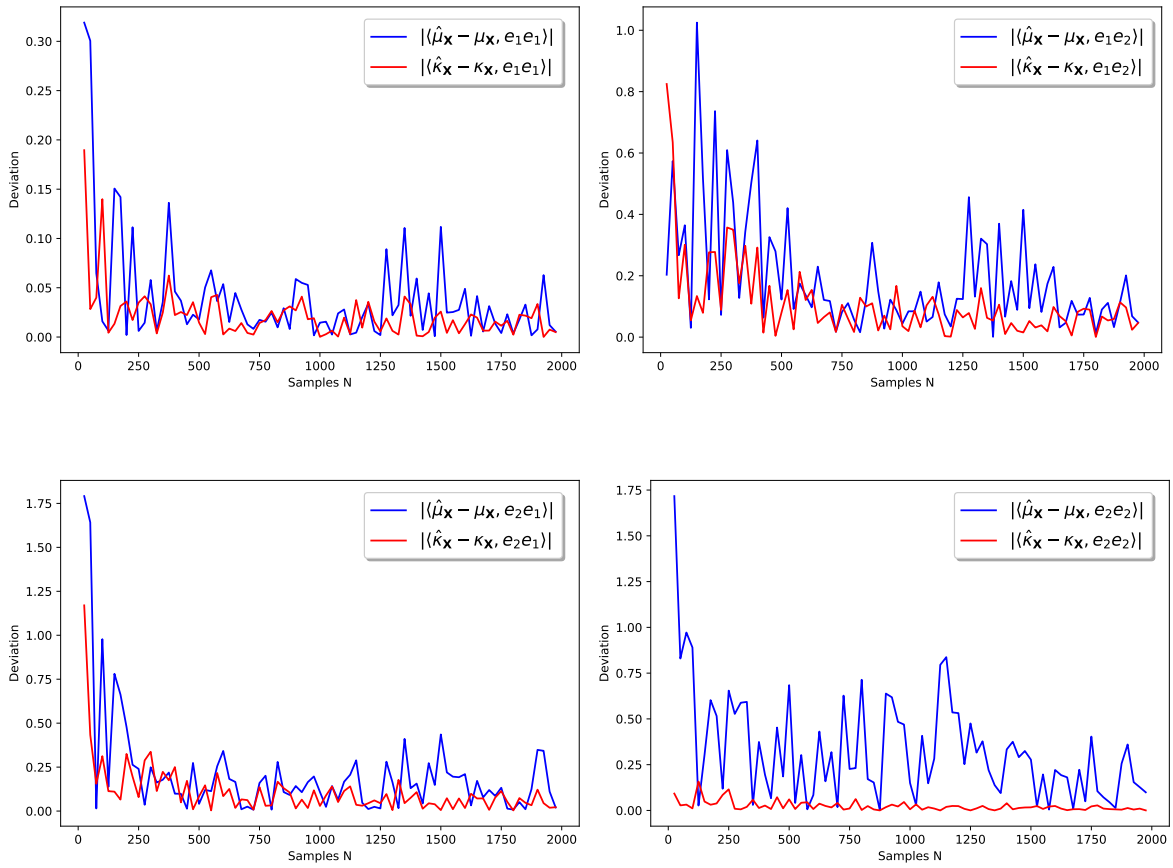


Figure 1.2: We simulated a 2-dimensional diffusion X with drift $b = \binom{1}{10}$ and volatility $\sigma = I_2$ the identity matrix. The plots show the absolute difference $|\hat{\mu}^2_{\mathbf{X}} - \mu^2_{\mathbf{X}}|$, and $|\hat{\kappa}^2_{\mathbf{X}} - \kappa^2_{\mathbf{X}}|$, between the theoretical values of the second signature moment (blue) and the second signature cumulant (red) and their estimators as the number of samples N varies in the range $25 \leq N \leq 2000$.

for $\kappa_X(\mathbf{a})$, $\hat{\kappa}_n(\tau)$ is a U-statistic for $\langle \kappa_{\mathbf{X}}, e_{\tau_1} ш \cdots ш e_{\tau_k} \rangle$ and this extends to any finite collection. It only remains to show the asymptotic variance. By Lemma 1.4.1 we get

$$\mathbb{E}\hat{\kappa}_n(\tau_1)\hat{\kappa}_n(\tau_2) - \mathbb{E}\hat{\kappa}_n(\tau_1)\mathbb{E}\hat{\kappa}_n(\tau_2) = \sum_{\substack{\mathbf{a}_1 \in \mathrm{Orp}(\tau_1) \\ \mathbf{a}_2 \in \mathrm{Orp}(\tau_2)}} \mathfrak{d}(\mathbf{a}_1)\mathfrak{d}(\mathbf{a}_2)\Big(\mathbb{E}[\hat{k}_n(\mathbf{a}_1)\hat{k}_n(\mathbf{a}_2)] - \kappa(\mathbf{a}_1)\kappa(\mathbf{a}_2)\Big)$$

$$= \sum_{\substack{\mathbf{a}_1 \in \mathrm{Orp}(\tau_1) \\ \mathbf{a}_2 \in \mathrm{Orp}(\tau_2)}} \mathfrak{d}(\mathbf{a}_1)\mathfrak{d}(\mathbf{a}_2)\mathrm{Cov}(\hat{k}_n(\mathbf{a}_1), \hat{k}_n(\mathbf{a}_2)) = \frac{1}{n} \sum_{\substack{\mathbf{a}_1 \in \mathrm{Orp}(\tau_1) \\ \mathbf{a}_2 \in \mathrm{Orp}(\tau_2)}} \mathfrak{d}(\mathbf{a}_1)\mathfrak{d}(\mathbf{a}_2)\mathcal{V}(\mathbf{a}_1, \mathbf{a}_2) + \mathcal{O}(n^{-2}).$$

□

Example 1.4.3 (Estimating a diffusion with constant drift and volatility)**.** Denote by B a standard d-dimensional Brownian motion and consider the process

$$X_t = bt + \sigma B_t,$$

where $b \in \mathbb{R}^d$ and $\sigma \in (\mathbb{R}^d)^{\otimes 2}$. It is well known that X can be lifted to a geometric rough path $\mathbf{X}$ and that

$$\mu_{\mathbf{X}} := \mathbb{E}[\mathbf{X}_{0,1}] = \exp(b + \frac{1}{2}\sigma)$$

see for instance [76, Exercise 3.22]. Hence, the first two signature cumulants and signature moments of $\mathbf{X}$ are

$$\begin{aligned} \mu^1_{\mathbf{X}} &= b, & \kappa^1_{\mathbf{X}} &= b, \\ \mu^2_{\mathbf{X}} &= \tfrac{1}{2}(b^{\otimes 2} + \sigma), & \kappa^2_{\mathbf{X}} &= \tfrac{1}{2}\sigma. \end{aligned}$$

By Item 2 of Theorem 1.3.6, both $\mu_{\mathbf{X}}$ and $\kappa_{\mathbf{X}}$ characterise the law of the process X, but if one wants to learn the law of the process X from observing sample trajectories of X, then it is much more efficient to work with signature cumulants than with signature moments, thus echoing a classic insight from statistics for vector-valued data, see Example 1.1.1. To see this, assume we are given an observation of the process $(X_t(\omega))_{t\in[0,N]}$ over a time interval $[0, N]$ for a large $N \in \mathbb{N}$. We divide this interval into N pieces of length 1 and calculate the signature of X over each of these shorter intervals. Since X is strong Markov, this results in N independent samples $\mathbf{X}_1, \dots, \mathbf{X}_N$ of the signature $\mathbf{X}$ of $(X_t)_{t\in[0,1]}$. The unbiased estimators introduced above are given as

$$\begin{aligned} \hat{\mu}^1_{\mathbf{X}} &= \tfrac{1}{N}\textstyle\sum_{i=1}^N \mathbf{X}^1_i, & \hat{\kappa}^1_{\mathbf{X}} &= \tfrac{1}{N}\textstyle\sum_{i=1}^N \mathbf{X}^1_i, \\ \hat{\mu}^2_{\mathbf{X}} &= \tfrac{1}{N}\textstyle\sum_{i=1}^N \mathbf{X}^2_i, & \hat{\kappa}^2_{\mathbf{X}} &= \tfrac{1}{N}\textstyle\sum_{i=1}^N \mathbf{X}^2_i - \tfrac{1}{2N(N-1)}\sum_{1\le i\neq j\le N} \mathbf{X}^1_i \otimes \mathbf{X}^1_j. \end{aligned}$$

We now show that the second signature moment typically has a higher variance than the second signature cumulant. In fact, a direct calculation detailed in Appendix 1.C shows that

$$\mathrm{Var}\Big(\langle \hat{\mu}_{\mathbf{X}}, e_i e_j\rangle\Big) = \mathrm{Var}\Big(\langle \hat{\kappa}_{\mathbf{X}}, e_i e_j\rangle\Big) + c_{ij}$$

where the last term c_{ij} is explicitly computed. For example, if $\sigma = I_d$ it reduces to

$$c_{ij} = \frac{1}{N}\begin{cases} \frac{1}{2}(b_i)^4 + 2(b_i)^2 - \frac{1}{2(N-1)}, & \text{if } i = j, \\ \frac{1}{2}(b_i)^2(b_j)^2 + \frac{1}{4}(b_i)^2 + \frac{1}{4}(b_j)^2 - \frac{1}{4(N-1)}, & \text{otherwise.} \end{cases}$$

and we see that the second signature cumulant estimator has a lower variance than the second signature moment estimator whenever either b or N is sufficiently large.

1.A Independence of Rough Paths on $\mathbb{R}^d$

For an element $x \in \mathrm{T}((V))$, let $\pi_n : \mathrm{T}((V)) \to V^{\otimes n}$ be the projection map and denote by $x^n = \pi_n x$. Further define:

$$E(V) = \{x \in \mathrm{T}((V)) : \sum_{n\geq 0} \|x^n\|\lambda^n < \infty, \, \forall \, \lambda > 0\},$$
$$G(V) = \{g \in E : \Delta(g) = g \otimes g, \, g \neq 0\}$$

where Δ is the map defined by extension of $\Delta(v) = v \otimes 1 + 1 \otimes v$ for $v \in V$. It follows from [129, Section 2.2.5] that signatures of weakly geometric rough paths in V take values in $G(V)$. Endow $E(V)$ with the locally convex topology induced by the family of norms $\|x\|_\lambda = \sum_{n\geq 0}\|x^n\|\lambda^n, \lambda > 0$ and $G(V)$ with the subspace topology. It follows from [44, Section 2 and 3] that:

1. $G(V)$ is Polish,
2. If $f \in E^\star$ there exists some $\lambda > 0$ such that $\|f \circ \pi_n\| \leq \lambda^n$ and $\sum_{n=0}^N f \circ \pi_n \to f$ pointwise,
3. If X is a random variable that takes values in $G(V)$ and $\mathbb{E}X \in E(V)$, then for $f \in E^\star$, $\mathbb{E}f(X) = \sum_{n\geq 0}\langle \mathbb{E}X^n, f\rangle$.

By [44, Section 4] there exists a subset of $E(V)^\star$ such that when restricted to $G(V)$ forms a $\star$-subalgebra of $C_b(G(V))$ that separates the points. Denote this family of functions by $\mathcal{C}(G)$.

We briefly recall the Stone-Weierstrass Theorem for Radon measures ([24, Exercise 7.14.79]): If $\mathcal{X}$ is a topological space and μ, ν are two Radon measures on $\mathcal{X}$. Then if $\mathcal{F} \subseteq C_b(\mathcal{X})$ is an algebra of functions that separates the points of $\mathcal{X}$, then $\mu = \nu$ is and only if $\mu(f) = \nu(f)$ for every $f \in \mathcal{F}$.

The next lemma is a slight generalisation of [44, Proposition 6.1] and its proof closely resembles that of [44, Proposition 3.2].

Lemma 1.A.1. *If $X_1, \dots, X_k$ are random variables taking values in $G(V)$ that satisfy:*

$$\sum_{n\geq 0} \|\mathbb{E}X_i^n\|\lambda^n < \infty, \quad \forall \lambda > 0$$

for $i = 1, \dots, k$. Then the joint law of $(X_1, \dots, X_k)$ is uniquely determined by:

$$\mathbb{E}\Big[\langle X_1, e_{\tau_1}\rangle \cdots \langle X_k, e_{\tau_k}\rangle\Big], \quad \tau_1, \dots, \tau_k \in [d]^\star.$$

Proof. The set of finite linear combinations of maps of the form

$$(x_1, \dots, x_k) \mapsto f_1(x_1) \cdots f_k(x_k), \quad f_1, \dots, f_k \in \mathcal{C}(G)$$

is a $\star$-subalgebra of $C_b(G^k)$ that separates the points of G^k. Since V is Polish, so is $G(V)$ and $G(V)^k$, so by [24, Theorem 7.1.7] the law of $(X_1, \dots, X_k)$ is Radon and hence by Stone-Weierestrass for Radon measures the assertion will then follow if we can show that Equation 1.A.1 determines $\mathbb{E}\Big[f_1(X_1) \cdots f_k(X_k)\Big]$ for every $f_1, \dots, f_k \in \mathcal{C}(G)$.

Fix some $f_1, \dots, f_k \in \mathcal{C}(G)$ and let $f_i^n = f_i \circ \pi_n$. For every $n \geq 0$, $f_i^n \in \mathrm{T}(V)$ so $\mathbb{E}f_i^n(X_i)$ is determined by Equation 1.A.1. Since we know that $\sum_{n=0}^N f_i^n \to f_i$ pointwise and $\mathbb{E}f_i(X_i) = \sum_{n\geq 0} \mathbb{E}f_i^n(X_i)$ it is enough to show that $\sum_{n_1,\dots,n_k \geq 0} \mathbb{E}(|f_1^{n_1}(X_1) \cdots f_k^{n_k}(X_k)|) < \infty$ since one may then apply dominated convergence to get:

$$\begin{aligned}\mathbb{E}\Big(f_1(X_1) \cdots f_k(X_k)\Big) &= \mathbb{E}\Big(\lim_{N\to\infty} \Big(\sum_{n_1=0}^N f_1^{n_1}(X_1)\Big) \cdots \Big(\sum_{n_k=0}^N f_k^{n_k}(X_k)\Big)\Big) \\ &= \sum_{n_1,\dots,n_k \geq 0} \mathbb{E}\Big(f_1^{n_1}(X_1) \cdots f_k^{n_k}(X_k)\Big).\end{aligned}$$

But for any $f \in E^*$ and $X \in G$ it holds that $f(X)^2 = f^{\otimes 2}(X \otimes X) = f^{\otimes 2}(\Delta X)$. Hence $f^2 = f^{\otimes 2} \circ \Delta \in E^*$ on G, so

$$\begin{aligned}&\sum_{n_1,\dots,n_k \geq 0} \mathbb{E}(|f_1^{n_1}(X_1) \cdots f_k^{n_k}(X_k)|) \leq \sum_{n_1,\dots,n_k \geq 0} \mathbb{E}|f_1^{n_1}(X_1)|^2 + \cdots + \mathbb{E}|f_1^{n_k}(X_k)|^2 \\ &= \sum_{n\geq 0} (f_1^n)^{\otimes 2}\Delta(\mathbb{E}X_1^{2n}) + \cdots + (f_k^n)^{\otimes 2}\Delta(\mathbb{E}X_k^{2n}).\end{aligned}$$

By continuity of $f_1, \dots, f_k, \Delta$ we may pick some $\lambda > 0$ such that $\|(f_i^n)^{\otimes 2} \circ \Delta\| \leq \lambda^{2n}$ for every $i = 1, \dots, k$. Now the assertion follows since

$$\sum_{n_1,\dots,n_k \geq 0} \mathbb{E}(|f_1^{n_1}(X_1) \cdots f_k^{n_k}(X_k)|) \leq \sum_{n\geq 0} \lambda^{2n}\Big[\|\mathbb{E}X_1^{2n}\| + \cdots + \|\mathbb{E}X_k^{2n}\|\Big] < \infty.$$

□

For the next Proposition, denote by $\mathcal{G}_p(V)$ the closure of C^1-paths in $\mathcal{G}_p^w(V)$. Even though $\mathcal{G}_p(V)$ is strictly smaller than $\mathcal{G}_p^w(V)$, one always has the inclusion $\mathcal{G}_p^w(V) \subseteq \mathcal{G}_q(V)$ if $q > p$ [76, Exercise 2.15]. Denote by $\mathcal{R}_p(V)$ the image of $\mathcal{G}_p(V)$ in $\mathcal{R}_p^w(V)$.

Proposition 1.A.2. *For any $p \geq 1$ and $\mathbf{X}, \mathbf{Y}$ taking values in $\mathcal{R}_p(V)$ such that*

$$\sum_{n\geq 0} \|\mathbb{E}\mathbf{X}_{0,T}^n\|\lambda^n < \infty, \quad \sum_{n\geq 0} \|\mathbb{E}\mathbf{Y}_{0,T}^n\|\lambda^n < \infty, \quad \forall\, \lambda > 0,$$

then $\mathbf{X}$ and $\mathbf{Y}$ are independent if and only if

$$\mathbb{E}\Big[\langle \mathbf{X}_{0,T}, e_{\tau_1}\rangle\langle \mathbf{Y}_{0,T}, e_{\tau_2}\rangle\Big] = \mathbb{E}\Big[\langle \mathbf{X}_{0,T}, e_{\tau_1}\rangle\Big]\mathbb{E}\Big[\langle \mathbf{Y}_{0,T}, e_{\tau_2}\rangle\Big] \quad \forall\, \tau_1, \tau_2 \in [d]^\star.$$

Proof. $\mathcal{G}_p(V)$ is separable [76, Exercise 2.12], hence so is $\mathcal{R}_p(V)$ as a continuous image of $\mathcal{G}_p(V)$. So the laws of $\mathbf{X}, \mathbf{Y}$ are radon [24, Theorem 7.1.7], and the map $\mathbf{X} \mapsto \mathbf{X}_{0,T}$ is a continuous injection [23, Theorem 1.1]. Since continuous injections into Hausdorff spaces are injective on Radon measures it is enough to show that $\mathbf{X}_{0,T}, \mathbf{Y}_{0,T}$ are independent.

But $\mathbf{X}_{0,T}$ and $\mathbf{Y}_{0,T}$ take values in $G(V)$, so the assertion follows by applying Lemma 1.A.1 to the measures on $G(V) \times G(V)$ defined by $\mu_1(\mathcal{A} \times \mathcal{B}) = \mathbb{P}(\mathbf{X}_{0,T} \in \mathcal{A}, \mathbf{Y}_{0,T} \in \mathcal{B}), \mu_2(\mathcal{A} \times \mathcal{B}) = \mathbb{P}(\mathbf{X}_{0,T} \in \mathcal{A})\mathbb{P}(\mathbf{Y}_{0,T} \in \mathcal{B})$ for Borel set $\mathcal{A}, \mathcal{B}$. □

1.B Tree-like equivalence of paths

For a path $x : [0,T] \to E$, where E is some topological space, denote by $\overleftarrow{x}$ its *time-reversal*, defined as

$$\overleftarrow{x}(t) := x(T-t).$$

For two paths $x : [0,T] \to E, y : [0,S] \to E$, denote by $x \star y : [0, T+S] \to E$ their *concatenation*, defined as

$$(x \star y)_t := \begin{cases} x(t) & \text{for} \quad 0 \le t < T, \\ y(s) + x(T) & \text{for} \quad T \le t \le T+S. \end{cases}$$

Definition 1.B.1 ([23])**.** A continuous path $x : [0,T] \to E$ is said to be *tree-like* if there exists an $\mathbb{R}$-tree τ, a continuous map $\varphi : [0,T] \to \tau$ and a map $\psi : \tau \to E$ such that $\varphi(0) = \varphi(T)$ and $x = \psi \circ \varphi$.

With the above definition in mind, we say that two paths x, y are *tree-like equivalent* if $x \star \overleftarrow{y}$ is tree-like. If $\mathbf{x}, \mathbf{y} \in \mathcal{G}_p^w(V)$ then we say that they are tree-like equivalent if the two paths $t \mapsto \mathbf{x}_{0,t}$ and $s \mapsto \mathbf{y}_{0,s}$ are. This induces an equivalence relation $\sim$ on $\mathcal{G}_p^w(V)$ and the corresponding quotient space

$$\mathcal{R}_p^w(V) = \mathcal{G}_p^w(V)/\sim$$

is called the space of *unparametrised* weakly geometric p-rough paths. This space is equipped with the topology induced by the map $\mathcal{R}_p^w(V) \to E(V)$, $\mathbf{x} \mapsto \mathbf{x}_{0,T}$.

Remark 1.B.1. Tree-like equivalence is essentially factoring out different ways of parametrising paths. If one wants to make a statement about a stochastic process X and not its unparametrised counterpart then one would simply look at the rough path lift of $(x_t, t)_{t \ge 0}$ instead since for any two paths x, y with the same starting value $(x_t, t) \sim (y_t, t)$ if and only if $x = y$.

Remark 1.B.2. Since a rough path $\mathbf{x}$ is a function of increments $(s,t) \mapsto \mathbf{x}_{s,t}$, the lift of a process x does not in general depend on its starting value x_0. If one wishes to make a statement about the whole process then one would lift (x_t, tx_0).

1.C Details for Example 1.4.3

For $i,j \in [d]$ denote with $\hat{\mu}_{\mathbf{X}}(i,j)$ the moment estimator associated to the partition $\mathbf{a} = i|j$ as seen in Example 1.4.1, then

$$\mathrm{Var}\Big(\langle \hat{\mu}_{\mathbf{X}}, e_i e_j\rangle\Big) = \mathrm{Var}\Big(\langle \hat{\kappa}_{\mathbf{X}}, e_i e_j\rangle\Big) - \frac{1}{4}\mathrm{Var}\Big(\hat{\mu}_{\mathbf{X}}(i,j)\Big) + \mathrm{Cov}\Big(\langle \hat{\mu}_{\mathbf{X}}, e_i e_j\rangle, \hat{\mu}_{\mathbf{X}}(i,j)\Big)$$

By using the formula in Remark 1.4.1 we can compute the formal expressions for the covariances of the polykays associated to the partitions $\mathbf{a} = a_1|a_2$ and $\mathbf{b} = a_3$,

$$\begin{aligned}
\mathrm{Cov}(\hat{k}(a_1|a_2), \hat{k}(a_3)) &= \frac{1}{N}\Big[\kappa(a_1|a_2a_3) + \kappa(a_2|a_1a_3) + 2\kappa(a_1|a_2|a_3)\Big] \\
\mathrm{Cov}(\hat{k}(a_1|a_2), \hat{k}(a_1|a_2)) &= \frac{1}{N}\Big[\kappa(a_1|a_1|a_2a_2) + \kappa(a_2|a_2|a_1a_1) + 2\kappa(a_1|a_2|a_1a_2)\Big] \\
&\quad + \frac{1}{N(N-1)}\Big[\kappa(a_1a_1|a_2a_2) + \kappa(a_1a_2|a_1a_2)\Big].
\end{aligned}$$

With this one sees that

$$\begin{aligned}
\mathrm{Cov}\Big(\langle \hat{\mu}_{\mathbf{X}}, e_i e_j\rangle, \hat{\mu}_{\mathbf{X}}(i,j)\Big) &= \frac{1}{2N}\Big(b_ib_j(\sigma_{ij}+\sigma_{ji}) + b_ib_i\sigma_{jj} + b_jb_j\sigma_{ii} + 2b_ib_j\sigma_{ij} + b_ib_ib_jb_j\Big) \\
\mathrm{Var}\Big(\hat{\mu}_{\mathbf{X}}(i,j)\Big) &= \frac{1}{N}\Big(b_ib_j(\sigma_{ij}+\sigma_{ji}) + b_ib_i\sigma_{jj} + b_jb_j\sigma_{ii}\Big) + \frac{1}{N(N-1)}\Big(\sigma_{ii}\sigma_{jj} + \sigma_{ij}\sigma_{ij}\Big),
\end{aligned}$$

and hence

$$c_{ij} := \frac{1}{4N}\Big(2b_ib_jb_jb_j + b_ib_j(\sigma_{ij}+\sigma_{ji}) + b_ib_i\sigma_{jj} + b_jb_j\sigma_{ii} + 4b_ib_j\sigma_{ij}\Big) - \frac{1}{4N(N-1)}\Big(\sigma_{ii}\sigma_{jj} + \sigma_{ij}\sigma_{ij}\Big).$$

1.D A generalisation to branched rough paths

We study the Hopf algebra of finite posets and give a formula for its iterated coproduct. This is used to provide a cancellation free formula for its antipode and also provides cancellation free moment-cumulants relations for its character group. These relations generalise and unify similar formulas for geometric, quasi geometric and branched rough paths.

1.D.1 The poset algebra

It is possible to define a commutative product and coproduct on posets. Given an order preserving function $f : P \to [n]$ we use the shorthand f_k for the poset $f^{-1}(k)$ with its induced ordering. We use the convention that if $f^{-1}(k) = \varnothing$ then $f_k = 1$.

Definition 1. Given two posets P_1, P_2 their commutative poset product P_1P_2 is defined as the set $P_1 \sqcup P_2$ with the partial order $x \leq y$ if $x, y \in P_i$ and $x \leq_{P_i} y$ and no further relations.

Definition 2. The poset coproduct Δ is defined by

$$\Delta(P) = \sum_{f \in \mathrm{Hom}(P,[2])} f_1 \otimes f_2.$$

where the sum is taken over all order preserving functions $f : P \to [2] = \{1, 2\}$. Equivalently this can be written as $\Delta(1) = 1^{\otimes 2}$ and for $P \neq 1$

$$\Delta(P) = P \otimes 1 + 1 \otimes P + \sum_{f:P \twoheadrightarrow [2]} f_1 \otimes f_2.$$

where the sum is taken over all surjective order preserving functions.

Remark 1.D.1. The coproduct can also be written in terms of ideals, see [2, Example 2.3]

Definition 3. Given a set $\mathcal{P}$ of finite posets that is closed under the poset product and coproduct, the poset algebra $\mathcal{H} = (\mathcal{P}, \cdot, \Delta)$ is the algebra of spanned by linear combinations of elements of P. $\mathcal{H}$ is graded by $|P|$ being the number of elements of P.

Proposition 1.D.1. *$\mathcal{H}$ is a Hopf algebra.*

Proof. Note that Δ is an algebra morphism since for any two posets F, G

$$\begin{aligned}\Delta(F)\Delta(G) &= (F \otimes 1 + 1 \otimes F + \sum_{f:F \twoheadrightarrow [2]} f_1 \otimes f_2)(G \otimes 1 + 1 \otimes G + \sum_{g:G \twoheadrightarrow [2]} g_1 \otimes g_2) \\ &= FG \otimes 1 + 1 \otimes FG + F \otimes G + G \otimes F + \sum_{g:G \twoheadrightarrow [2], f:F \twoheadrightarrow [2]} f_1 g_1 \otimes f_2 g_2.\end{aligned}$$

We note that $\mathrm{Hom}(FG, [2])$ can be decomposed as either $f^{-1}(x) = G, f^{-1}(y) = F$ for $x, y = 1, 2$ or $f^{-1}(1) = A \cup B, f^{-1}(2) = C \cup D$ where A, C is and ordered partition of F and B, D is an ordered partition of G. Hence

$$F \otimes G + G \otimes F + \sum_{g:G \twoheadrightarrow [2], f:F \twoheadrightarrow [2]} f_1 g_1 \otimes f_2 g_2 = \sum_{f:FG \twoheadrightarrow [2]} f_1 \otimes f_2$$

which implies that $\Delta(F)\Delta(G) = \Delta(FG)$. □

Proposition 1.D.2. *Given a poset $P \in \mathcal{H}$, it holds that*

$$\Delta^{n-1}P = \sum_{f \in \mathrm{Hom}(P,[n])} f_1 \otimes \cdots \otimes f_n$$

where the sum is taken over all order preserving maps $f : P \to [n]$.

Proof. This is shown by induction on n, for $n = 2$ this is clear by definition. Assume that it holds for $n-1$ and write

$$\begin{aligned}\Delta^n P =& (\Delta \otimes 1^{\otimes n-1}) \sum_{f \in \mathrm{Hom}(P,[n])} f_1 \otimes \cdots \otimes f_n \\ =& \sum_{f \in \mathrm{Hom}(P,[n])} \big(\Delta f_1\big) \otimes \cdots \otimes f_n \\ =& \sum_{f \in \mathrm{Hom}(P,[n])} \sum_{g \in \mathrm{Hom}(f_1,[2])} g_1 \otimes g_2 \otimes f_2 \otimes \cdots \otimes f_n\end{aligned}$$

and the assertion follows from noting that every term in $\mathrm{Hom}(P,[n+1])$ will appear exactly once in the above sum. To see this, note that for any $h \in \mathrm{Hom}(P,[n+1])$ we can define f, g as follows $g_1 = h_1, g_2 = h_2$ and $f_1 = h_1 \cup h_2$ and $f_k = h_{k+1}$ for $k > 1$, this shows that every $h \in \mathrm{Hom}(P,[n+1])$ appears at least once in the sum. To see that they appear at most once, note that if $g_1 \otimes g_2 \otimes f_2 \otimes \cdots \otimes f_n = g_1' \otimes g_2' \otimes f_2' \otimes \cdots \otimes f_n'$ then clearly $g = g'$ which implies that $f = f'$. □

Remark 1.D.2. Note that by restricting to $(\mathcal{H}_0^\star)^{\otimes n}$ we get

$$\Delta^{n-1}P|_{(\mathcal{H}_0^\star)^{\otimes n}} = \sum_{f : F \twoheadrightarrow [n]} f_1 \otimes \cdots \otimes f_n.$$

where the sum is taken over all surjective order preserving maps.

Example 1.D.1.

$$\begin{aligned}\Delta^2 P = P \otimes 1 + 1 \otimes P + \sum_{f : P \twoheadrightarrow [2]} f_1 \otimes f_2 \otimes 1 + f_1 \otimes 1 \otimes f_2 + 1 \otimes f_1 \otimes f_2 \\ + \sum_{f : P \twoheadrightarrow [3]} f_1 \otimes f_2 \otimes f_3\end{aligned}$$

Recall the following definitions from the main text:

- The ordered partitions of a poset P is defined as

$$\mathrm{OP}(P) = \{\ker . f : f \in \mathrm{Hom}(P, \mathbb{N})\}$$

- Given a partially ordered partition π, its factorial is defined as

$$\pi! = \#\{f \in \mathrm{Hom}(P, [|P|]) : \ker . f = \pi\}$$

- For a given partially ordered partition π on P, we define its *antichain ancestry* as

$$A(\pi) = \{b \geq \pi : b \cap C = \pi \cap C \text{ for any chain } C \subseteq P\}.$$

  Note that if $b \in A(\pi)$, then $\pi_1 \cdots \pi_{|\pi|} = b_1 \cdots b_{|b|}$.

Two elements x, y of a poset P are said to be *connected* if there exists a sequence of elements $z_1, \ldots, z_k \in P$ such that $z_1 \leq z_{i+1}$ or $z_{i+1} \leq z_i$ and $z_1 = x, z_k = y$. A poset is said to be connected if any two elements are. A partition is said to be connected if all of its blocks are connected. The set of connected partially ordered partitions of a poset P is denoted by $C(P)$.

Lemma 1.D.3. *$C(P)$ is exactly the set of ordered partitions that are not contained in any antichain ancestry except their own.*

Proof. Pick $\pi \in C(P)$ and assume that $\pi \in A(\Pi)$. Then any block of π can be written as a union of blocks of Π by assumption. Pick any two elements $x, y \in P$ that belong to the same block $\pi^\star \in \pi$ and let $\Pi^\star$ be the block of Π that contains x. By assumption there exists a sequence $C_1, \ldots, C_k \subseteq \pi^\star$ of intersecting chains such that $x \in C_1, y \in C_k$. Since $\pi \cap C_i = \Pi \cap C_i$ for $1 \leq i \leq k$ it must hold that $C_i \subseteq \Pi^\star$ for every i, hence $y \in \Pi^\star$ and it must hold that $\pi^\star = \Pi^\star$. So $\Pi = \pi$. □

The forest algebra

Definition 4. Given a poset P and an element $w \in P$, the ideal $I_w \subseteq P$ is defined as $\{v \in P : v \leq w\}$. We say that a poset P is a *forest* if I_w is totally ordered for any $w \in P$. We say that P is a rooted tree if it has a unique smallest element. Note that any forest can be written as a disjoint union of trees. Note that any subset of a forest is also a forest with its induced order.

Remark 1.D.3. A forest can also be defined as a graph where every pair of vertices are either connected by a unique path or not at all. This is equivalent to the definition above by the following construction: Given a vertex $v \in F$ in a forest F, let γ_v be the set of vertices visited by the shortest path from v to its root. We define a partial order on F by setting

$$w \leq v \text{ if } w \in \gamma_v$$

then F is a forest per the above definition.

A *decorated forest* is a pair (F, h) where F is a forest and $h : F \to [d]$ is a function that assigns a value from 1 to d to every node in F.

Proposition 1.D.4. *The Connes-Kreimer Hopf algebra is the poset algebra on decorated forests.*

Proof. For a tree τ, any order preserving $f : \tau \to [2]$ defines an *admissible cut* on τ by definition, hence the coproduct agrees with the Connes Kreimer coproduct on trees, so they are the same. □

Proposition 1.D.5. *Any connected partition of a forest is ordered*

Proof. Let π be a connected partition of a forest F. Since π is connected it can be written as $\pi = (\pi \cap F_1) \sqcup \cdots (\pi \cap F_k)$ where $F_1 \cdots F_k = F$ are the trees that make up F. Since the trees are incomparable we may assume without loss of generality that F is a tree. Now pick the block of π that contains the root, call it π_1. Note that for any $x \in \pi_1$ and $y \notin \pi_1$ it must hold that either they are incomparable, or $x \leq y$ since if $y \leq x$ then since I_x is totally ordered and π_1 is connected $y \in \pi_1$. Hence $F \backslash \pi_1$ is a forest with at most $|F| - 1$ elements, and by induction we may construct an order preserving function that induces π by setting $f(1) = \pi_1$ and repeating the procedure. □

Proposition 1.D.6. *If F is a tree τ, then a connected ordered partition π inherits a tree structure from τ by connecting the successive blocks of π. $\pi!$ can then be defined inductively by*

$$\cdot! = 1, \quad [\pi_1, \ldots, \pi_k]! = \binom{|\pi_1| + \cdots + |\pi_k|}{|\pi_1|, \ldots, |\pi_k|} \pi_1! \cdots \pi_k!$$

Remark 1.D.4. If F is a proper forest then an ordered partition π will not be a tree or a forest in general and can not be computed iteratively in the same way. However, if π satisfies that none if its blocks contains elements from more than one tree, then it does inherit a forest structure and $\pi!$ can be computed as

$$\pi! = \frac{n!(\pi \cap \tau_1)! \cdots (\pi \cap \tau_n)!}{|\pi \cap \tau_1|! \cdots |\pi \cap \tau_n|!}$$

The chain algebra

Just like the forest algebra is the algebra generated by trees, the chain algebra is the algebra generated by chains

Definition 5. The chain algebra is the poset algebra on disjoint unions of decorated chains.

Remark 1.D.5. Since a union of chains is a forest, the chain algebra is a subalgebra of the Connes Kreimer algebra.

The free symmetric algebra

The free symmetric algebra of dimension d is the poset algebra on d posets with only one element each, note that in this case, any function is order preserving, hence we may write the coproduct of any element in the algebra as

$$\Delta(p_1^{n_1}\cdots p_k^{n_k}) = \sum_{I\sqcup J=[|P|]} p_1^{i_1}\cdots p_k^{i_k}\otimes p_1^{j_1}\cdots p_k^{j_k}$$

1.D.2 Moment cumulant relations

Definition 6. The log and exp functions are defined on $\mathcal{H}^\star$ by

$$\log(x) = \sum_{n\geq 1}\frac{(-1)^{n-1}}{n}(x-1)^n$$
$$\exp(x) = \sum_{n\geq 0}\frac{1}{n!}x^n$$

where the product is the multiplication on $\mathcal{H}^\star$.

Given a random character X on $\mathcal{H}^\star$ we define its moments as $\mathbb{E}X$ and its cumulants as $\log \mathbb{E}X$.

Proposition 1.D.7. *For any poset $P\in\mathcal{H}$ and random character X on $\mathcal{H}^\star$ with cumulants κ and moments μ it holds that*

$$\langle P,\kappa\rangle = \sum_{\pi\in OP(P)}(-1)^{|\pi|-1}\frac{\pi!}{|\pi|}\mu(\pi) = \sum_{\pi\in OP(P)}(-1)^{n-1}\frac{\pi!}{n}\langle \pi_1\otimes\cdots\otimes\pi_n,\mu^{\otimes n}\rangle$$

where π runs over all ordered partitions of P.

Proof. Note that

$$\langle P,\kappa\rangle = \langle P,\log\mathbb{E}X\rangle = \sum_{n\geq 1}\frac{(-1)^{n-1}}{n}\langle P,(\mathbb{E}X-1)^n\rangle = \sum_{n\geq 1}\frac{(-1)^{n-1}}{n}\langle \Delta^n P,(\mathbb{E}X-1)^{\otimes n}\rangle$$

so it is enough to show that for any $n>0$

$$\langle \Delta^n P,(\mathbb{E}X-1)^{\otimes n}\rangle = \sum_{\pi\in \mathrm{OP}(P):|\pi|=n}\pi!\mu(\pi) = \sum_{f\twoheadrightarrow[n]}\mu(\ker . f) = \sum_{f\twoheadrightarrow[n]}\langle f_1,\mathbb{E}X\rangle\cdots\langle f_n,\mathbb{E}X\rangle.$$

which follows from Lemma 1.D.2. □

Remark 1.D.6. Note that Proposition 1.D.7 is more general than it appears since we have the sequence of inclusions

$$\text{Chain algebra} \subseteq \text{Connes-Kreimer algebra} \subseteq \text{Poset algebra}.$$

Therefore it generalises the results of [30], which studies geometric rough paths, to a setting that includes branched rough paths as well as quasi-geometric rough paths

Corollary 1.D.8. *If X is a branched rough path on $\mathbb{R}^d$ with signature $S(X)$ such that $\mathbb{E}S(X)$ has finite coordinate values, then it holds that for any forest F*

$$\langle F, \log \mathbb{E}S(X)\rangle = \sum_{\pi \in OP(F)} (-1)^{|\pi|-1} \frac{\pi!}{|\pi|} \langle \pi_1, \mathbb{E}S(X)\rangle \cdots \langle \pi_{|\pi|}, \mathbb{E}S(X)\rangle$$

Remark 1.D.7. In the symmetric algebra, since any function is order preserving we get that for any partially ordered partition $\pi! = |\pi|!$. This implies that

$$\log \mathbb{E}p_1^{n_1} \cdots p_k^{n_k} = \sum_{\pi} (-1)^{|\pi|-1}(|\pi|-1)!\mathbb{E}(p_1^{\pi_1^1} \cdots p_k^{\pi_k^1}) \cdots \mathbb{E}(p_1^{\pi_1^{|\pi|}} \cdots p_k^{\pi_k^{|\pi|}})$$

which are the usual moment cumulant relations on vector valued data.

Remark 1.D.8. If C is a single chain in the chain algebra, then the ordered partitions and order preserving functions are in 1 to 1 correspondence, hence we get

$$\langle C, \kappa\rangle = \sum_{\pi \in \mathrm{OP}(C)} (-1)^{|\pi|-1} \mu(\pi_1) \cdots \mu(\pi_{|\pi|})$$

where the sum is taken over all partitions with convex blocks.

Chapter 2

Kernelized Cumulants: Beyond Kernel Mean Embeddings

This chapter is based on [31] which was written with Harald Oberhauser and Zoltán Szabó.

In $\mathbb{R}^d$, it is well-known that cumulants provide an alternative to moments that can achieve the same goals with numerous benefits such as lower variance estimators. In this paper we extend cumulants to reproducing kernel Hilbert spaces (RKHS) using tools from tensor algebras and show that they are computationally tractable by a kernel trick. These kernelized cumulants provide a new set of all-purpose statistics; the classical maximum mean discrepancy and Hilbert-Schmidt independence criterion arise as the degree one objects in our general construction. We argue both theoretically and empirically (on synthetic, environmental, and traffic data analysis) that going beyond degree one has several advantages and can be achieved with the same computational complexity and minimal overhead in our experiments.

2.1 Introduction

The moments of a random variable in $\mathbb{R}^d$ are arguably the most popular all-purpose statistic. However, for many purposes, cumulants [133] are much better suited statistics than moments. For example, low-degree moments can dominate higher degree moments which becomes particular concerning when only finite samples are available. As an elementary example, we note that the unbiased estimator for the second cumulant generally has a lower variance than the unbiased estimator for the second moment: denoting the minimum variance unbiased estimators by $\widehat{\mu^2}$ and $\widehat{\kappa}$ respectively, one can show that given N samples from a random variable X, the following holds

$$\operatorname{Var}\left(\widehat{\mu^2}\right) = \operatorname{Var}\left(\widehat{\kappa}\right) + \frac{2}{N}\left[(\mathbb{E}X)^4 - (\mathbb{E}X)^2 \operatorname{Var}(X) - 2\frac{\operatorname{Var}(X)^2}{N-1}\right];$$

see Appendix 2.A for details. Similarly, for independence testing cumulants are classical since independence amounts to testing if the cross-cumulants vanishing. Most people will be familiar with the fact that the covariance – the second cumulant (the first being the mean) – is zero for independent random variables but this property is not sufficient for independence (all cross-cumulants must vanish). Further, cumulants have many other attractive properties such as a simple characterization of the Gaussian distribution and additivity over independent random variables which often makes them better suited than moments when working for instance with i.i.d. random variables or Gaussians. In this paper, we deploy (reproducing) kernels to study the distribution of a random variable by lifting it into a RKHS. The contribution of this article is to go beyond the mean in the RKHS by developing the notion of cumulants in RKHSs. We show that probably the most fundamental of the above advantages of classical cumulants of $\mathbb{R}^d$-valued random variables persist; a kernel trick yields efficient computation of such cumulant-based statistics and we apply these insight to experiments about two-sample and independence testing and achieve improved test power.

Kernel embeddings. Mean and covariance arise naturally in the context of kernel-enriched domains. Kernel techniques [164, 177, 157] provide a principled and powerful approach for lifting data points to a so-called reproducing kernel Hilbert space (RKHS; [5]). Considering the mean of this feature – referred to as kernel mean embedding (KME; [17, 170]) – also enables one to represent probability measures, and to induce a semi-metric referred to as maximum mean discrepancy (MMD; [170, 92]) and an

independence measure called the Hilbert-Schmidt independence criterion (HSIC[1]; [93]). It is known that when the underlying kernel is *characteristic* [80, 174], the MMD is a metric. HSIC captures independence for 2 components with characteristic kernels and for > 2 components [151, 166] with universal ones [183]. MMD belongs to the family of integral probability metrics (IPM; [208, 137]) when in the IPM the underlying function class is chosen to be the unit ball of the RKHS. MMD and HSIC are known to be equivalent [167] to the notions of energy distance [9, 184, 185]—also called N-distance [207, 114]– and energy distance [187, 186, 128] of the statistics literature. Both MMD and HSIC can be expressed in terms of expectations of kernel values which can be leveraged to design efficient estimators for them; we will refer to this trick as the *expected kernel* trick. A recent survey on mean embedding and their applications is given by [136].

Kernelized cumulants. This paper aims to extend the notion of cumulants to RKHSs and combine the favorable properties of the two fields. The primary technical challenge one has to resolve is that cumulants have a quite rich algebraic structure that is closely linked to the partition lattice [171, 172], and requires tools from algebra and combinatorics to be studied. The key idea of the extension is that the classical defining property of cumulants of random variables in $\mathbb{R}^d$ via the logarithm of the moment generating function can be generalized by using tools from tensor algebras. Although cumulants are classic in probability they do not seem to have received attention in the kernel literature; partly a reason might be the underlying algebraic structure is already non-trivial in $\mathbb{R}^d$; for example, even extending their definition to RKHS needs care since this relies on the algebraic structure of series of tensors in RKHSs. The closest previous work to our knowledge is by [132] and considers the variance in the RKHS – which in our setting can be identified as the first kernelized cumulant after the kernel mean embedding – for two-sample testing like we do in parts of this paper. But this is never put in the context of cumulant embeddings, nor their theoretical guarantees are derived. Our main theoretical contribution is to provide a framework for general cumulants (which includes but also goes beyond variance) in RKHS and to provide sufficient conditions (Theorem 2.3.1 and Theorem 2.3.2) on the kernel itself that theoretically justify two-sample and independence testing that applies to a large class of kernels which we term point-separating.

[1]HSIC is a MMD with the tensor product kernel evaluated on the joint distribution and the product of the marginals, or equivalently the Hilbert-Schmidt norm of the cross-covariance operator.

Outline. The paper is structured as follows: In Section 2.2 we formulate the notion of cumulants of random variables in Hilbert spaces. In Section 2.3 we prove a kernelized version of Theorem 2.2.1 that generalizes a classic result of $\mathbb{R}^d$-valued random variables about the characterization of distributions using cumulants. We go on to show that for RKHSs, we may leverage the expected kernel trick to derive efficient estimators for novel statistics; KME and HSIC arise as the "degree 1" objects of these statistics. In Section 2.4 we demonstrate numerically that going beyond degree 1 is advantageous; even just going to degree 2 can lead to significant improvement and comes with little computational overhead. We provide general technical background (on cumulants, tensor products, and tensor sums of Hilbert spaces and tensor algebras), proofs, further details on numerical experiments, and our V-statistic based estimators in the Appendices.

2.2 Moments and Cumulants

We briefly revisit classical cumulants and define cumulants of random variables in Hilbert spaces.

Moments. Let γ be a probability measure on $\mathbb{R}^d$ and $(X_1, \ldots, X_d) \sim \gamma$. The moments $\mu(\gamma) = (\mu^{\mathbf{i}}(\gamma))_{\mathbf{i}\in\mathbb{N}^d}$ of γ are defined as

$$\mu^{\mathbf{i}}(\gamma) := \mathbb{E}\left[X_1^{i_1} \cdots X_d^{i_d}\right] \in \mathbb{R}, \tag{2.1}$$

where $\mathbf{i} = (i_1, \ldots, i_d) \in \mathbb{N}^d$ denotes a d-tuple of non-negative integers $(i_1, \ldots, i_d \geq 0)$. The *degree* of an element $\mathbf{i} \in \mathbb{N}^d$ is defined as $\deg(\mathbf{i}) := i_1 + \cdots + i_d$. For $m \in \mathbb{N}$, let $\mu^m(\gamma) := (\mu^{\mathbf{i}}(\gamma))_{\deg(\mathbf{i})=m}$ which we refer to as the m-th moments of γ with the convention that $\mu^0(\gamma) = 1$.

Cumulants. Cumulants $\kappa(\gamma) = (\kappa^{\mathbf{i}}(\gamma))_{\mathbf{i}\in\mathbb{N}^d}$ can be defined by the moment generating function as

$$\sum_{\mathbf{i}\in\mathbb{N}^d} \kappa^{\mathbf{i}}(\gamma)\frac{\theta^{\mathbf{i}}}{\mathbf{i}!} = \log \sum_{\mathbf{i}\in\mathbb{N}^d} \mu^{\mathbf{i}}(\gamma)\frac{\theta^{\mathbf{i}}}{\mathbf{i}!},\ \theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d, \tag{2.2}$$

where we denote $\mathbf{i}! = i_1! \cdots i_d!$ and $\theta^{\mathbf{i}} = \theta_1^{i_1} \cdots \theta_d^{i_d}$; an equivalent definition of cumulants is via a combinatorial expression of partitions (elaborated in Appendix 2.C.1). Cumulants have several attractive properties, the following forms our main motivation.

Theorem 2.2.1 (Characterization of distributions with cumulants on $\mathbb{R}^d$, [20])**.** *Let γ be a probability measure on a bounded subset of $\mathbb{R}^d$ with cumulants $\kappa(\gamma)$ and let $(X_1, \ldots, X_d) \sim \gamma$. Then*

1. *$\gamma \mapsto \kappa(\gamma)$ is injective.*
2. *$X_1, \ldots, X_d$ are jointly independent if and only if $\kappa^{\mathbf{i}}(\gamma) = 0$ for all d-tuples of positive integers $\mathbf{i} \in \mathbb{N}_+^d$.*

2.2.1 Moments in Hilbert Spaces

Instead of directly considering the law of a tuple of random variables $(X_1, \ldots, X_d)$ in a product space $\mathcal{X}_1 \times \cdots \times \mathcal{X}_d$, it can be advantageous to use feature maps $\Phi_i : \mathcal{X}_i \to \mathcal{H}_i$ and instead study the distribution of the $\mathcal{H}_1 \times \cdots \times \mathcal{H}_d$-valued random variable $\big(\Phi_1(X_1), \ldots, \Phi_d(X_d)\big)$. Motivated by this lifting, we study here moments of Hilbert-space valued random variables and assume in this subsection (with a slight abuse of notations) that one has already applied the lifting and $X_i \in \mathcal{H}_i$ where $i = 1, \ldots, d$. In Section 2.3 we specialize the construction to RKHSs, and use these moments (Def. 7) to define kernelized cumulants.

Moments. In the finite-dimensional case (2.2) we defined the moment sequence by taking expectations of products of the coordinates of the underlying random variable. For the infinite-dimensional case, it is convenient to develop a coordinate-free definition which can be accomplished by using tensors. To do so we make use of the following results about Hilbert spaces: for real Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ the tensor product $\mathcal{H}_1 \otimes \mathcal{H}_2$ is the Hilbert space given by completion of the tensor product of $\mathcal{H}_1$ and $\mathcal{H}_2$ as vector space; we also write $\mathcal{H}_1^{\otimes m} := \underbrace{\mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_1}_{\text{m-times}}$. Similarly, the direct sum $\mathcal{H}_1 \oplus \mathcal{H}_2$ is a Hilbert space. It is natural to consider $\mathbb{E}\left[X_1^{\otimes m}\right] \in \mathcal{H}_1^{\otimes m}$ as the m-th moment of a $\mathcal{H}_1$-valued random variable X_1 where the integral in the expectation is meant in Bochner sense. Consequently the natural state space for all moments of a $\mathcal{H}_1$-valued random variable is the tensor algebra $\mathrm{T}_1 := \prod_{m \geq 0} \mathcal{H}_1^{\otimes m}$ where by convention $\mathcal{H}_1^{\otimes 0} := \mathbb{R}$. See Appendix 2.B for more details on tensor products of Hilbert spaces and tensor algebras.

Example 2.2.1 ($\mathcal{H}_1 = \mathbb{R}^d$, $m = 2$)**.** If $X_1 = \left(X_1^1, \ldots, X_1^d\right)$ is $\mathcal{H}_1 = \mathbb{R}^d$-valued then $\mathbb{E}\left[X_1^{\otimes 2}\right] \in (\mathbb{R}^d)^{\otimes 2}$ can be identified with a $(d \times d)$-sized matrix whose (i, j)-th entry is $\mathbb{E}\left[X_1^i X_1^j\right]$.

Since we are interested in the general case of a $\mathcal{H}_1 \times \cdots \times \mathcal{H}_d$-valued random variable $X = (X_1, \ldots, X_d)$ we arrive at the definition below.

Definition 7 (Moments in Hilbert spaces)**.** Let γ be a probability measure on $\mathcal{H} := \mathcal{H}_1 \times \cdots \times \mathcal{H}_d$ and let $(X_1, \ldots, X_d) \sim \gamma$. We define

$$\mu^{\mathbf{i}}(\gamma) := \mathbb{E}[X_1^{\otimes i_1} \otimes \cdots \otimes X_d^{\otimes i_d}] \in \mathcal{H}^{\otimes \mathbf{i}}, \quad \mathcal{H}^{\otimes \mathbf{i}} := \mathcal{H}_1^{\otimes i_1} \otimes \cdots \otimes \mathcal{H}_d^{\otimes i_d} \tag{2.3}$$

for every $\mathbf{i} \in \mathbb{N}^d$ whenever the above expectation exists. The moment sequence is defined as the element

$$\mu(\gamma) = (\mu^{\mathbf{i}}(\gamma))_{\mathbf{i} \in \mathbb{N}^d} \in \mathrm{T} := \mathrm{T}_1 \otimes \cdots \otimes \mathrm{T}_d, \text{ with } \mathrm{T}_j := \prod_{m \geq 0} \mathcal{H}_j^{\otimes m},$$

and for $m \in \mathbb{N}$ we refer to $\mu^m(\gamma) = \bigoplus_{\mathbf{i} \in \mathbb{N}^d : \deg(\mathbf{i}) = m} \mu^{\mathbf{i}}(\gamma)$ as the m-moments of γ.

In case of $\mathcal{H}_i = \mathbb{R}$, both definitions (2.1) and (2.3) apply for $\mu^{\mathbf{i}}(\gamma)$. Henceforth, we always refer to (2.3) when we write $\mu^{\mathbf{i}}(\gamma)$. Even in the finite-dimensional case, Def. 7 is useful, for instance when $X_1 \in \mathcal{H}_1$ and $X_2 \in \mathcal{H}_2$ have different state space ($\mathcal{H}_1 \neq \mathcal{H}_2$).

2.3 Kernelized Cumulants

We lift a random variable $X = (X_1, \ldots, X_d) \in \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ via a feature map $\Phi : \mathcal{X} \to \mathcal{H}$ into a Hilbert space valued random variable $\Phi(X)$. For the rest of the paper (i) $\mathcal{X}_1, \ldots, \mathcal{X}_d$ will denote a collection of Polish spaces, but the reader is invited to think of them as finite-dimensional Euclidean spaces, (ii) $\mathcal{H}$ is an RKHS with kernel k and canonical feature map $\Phi(x) = k(x, \cdot)$,[2] and (iii) all kernels are assumed to be bounded.[3] Our main results (Theorem 2.3.1 and Theorem 2.3.2) are that in this case the expected kernel trick applies to both items in the kernelized version of Theorem 2.2.1. The key to these results is an expression for inner products of cumulants in RKHSs (Lemma 2.3.1).

A Combinatorial Expression of Cumulants. Classical cumulants can be defined via the moment generating function or via combinatorial sums over *partitions* (Appendix 2.C.1). To generalize cumulants to RKHSs the combinatorial definition is the most efficient way. A partition π of m elements is a family of non-empty, disjoint subsets $\pi_1, \ldots, \pi_b$ of

[2] $k(x, \cdot)$ denotes the function $x' \mapsto k(x, x')$ with $x \in \mathcal{X}$ fixed.

[3] A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called bounded if there exists $B \in \mathbb{R}$ such that $\sup_{x, x' \in \mathcal{X}} k(x, x') \leq B$.

$\{1, \ldots, m\}$ whose union is the whole set; formally $\bigcup_{j=1}^{b} \pi_j = \{1, \ldots, m\}$ and $\pi_i \cap \pi_j = \varnothing$ for $i \neq j$. We call b the number of blocks of the partition π and use the shorthand $|\pi|$ to denote it. The set of all partitions of m is denoted with $P(m)$. To formulate our main results, it is convenient to associate with a measure γ and a partition π the so-called partition measure γ_π that is given by permuting the marginals of γ.

Definition 8 (Partition measure)**.** Let γ be a probability measure on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ and $\pi \in P(d)$. Define

$$\gamma_\pi := \gamma|_{\mathcal{X}_{\pi_1}} \otimes \cdots \otimes \gamma|_{\mathcal{X}_{\pi_b}},$$

where $\mathcal{X}_{\pi_i}$ denotes the product space $\prod_{j \in \pi_i} \mathcal{X}_j$ and $\gamma|_{\mathcal{X}_{\pi_i}}$ is the corresponding marginal distribution of γ. We call γ_π the partition measure induced by π.

We also associate with γ and a multi-index $\mathbf{i}$ the so-called diagonal measure $\gamma^{\mathbf{i}}$ that is given by repeating marginals according to $\mathbf{i}$.

Definition 9 (Diagonal measure)**.** Let γ be a probability measure on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ and $\mathbf{i} = (i_1, \ldots, i_d) \in \mathbb{N}^d$. Define

$$\gamma^{\mathbf{i}} := \mathrm{Law}(\underbrace{X_1, \ldots, X_1}_{i_1 \text{ times}}, \underbrace{X_2, \ldots, X_2}_{i_2 \text{ times}}, \ldots, \underbrace{X_d, \ldots, X_d}_{i_d \text{ times}}),$$

where $(X_1, \ldots, X_d) \sim \gamma$. We call $\gamma^{\mathbf{i}}$ the diagonal measure induced by $\mathbf{i}$.

In general, the partition measure γ_π and the diagonal measure are not probability measures on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ but on spaces that are constructed by permuting or repeating $\mathcal{X}_1, \ldots, \mathcal{X}_d$. Formally, γ_π is a probability measure on $\mathcal{X}_{\pi_1} \times \cdots \times \mathcal{X}_{\pi_b}$ and $\gamma^{\mathbf{i}}$ is a probability measure on $\mathcal{X}_1^{i_1} \times \cdots \times \mathcal{X}_d^{i_d}$; thus, γ_π has d coordinates and $\gamma^{\mathbf{i}}$ has $\deg(\mathbf{i})$ coordinates. These two constructions can be combined, writing $\gamma_\pi^{\mathbf{i}}$ for the measure $(\gamma^{\mathbf{i}})_\pi$ which makes sense whenever $\pi \in P(\deg(\mathbf{i}))$. We can now write down our generalization of cumulants.

Definition 10 (Kernelized cumulants)**.** Let γ be a probability measure on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ and let $(\mathcal{H}_1, k_1), \ldots, (\mathcal{H}_d, k_d)$ be RKHSs on $\mathcal{X}_1, \ldots, \mathcal{X}_d$ respectively. We define the kernelized cumulants

$$\kappa_{k_1,\ldots,k_d}(\gamma) := \left(\kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma)\right)_{\mathbf{i} \in \mathbb{N}^d} \in \mathrm{T}$$

as follows

$$\kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma) := \sum_{\pi \in P(m)} c_\pi \mathbb{E}_{\gamma_\pi^{\mathbf{i}}} k^{\otimes \mathbf{i}}((X_1, \ldots, X_m), \cdot),$$

where $m = \deg(\mathbf{i})$, $c_\pi := (-1)^{|\pi|-1}(|\pi|-1)!$, $\gamma_\pi^{\mathbf{i}} = (\gamma^{\mathbf{i}})_\pi$ and

$$k^{\otimes \mathbf{i}}((x_1,\ldots,x_m),(y_1,\ldots,y_m)) := k_1(x_1,y_1)\cdots k_1(x_{i_1},y_{i_1}) \\ \cdots k_d(x_{m-i_d+1},y_{m-i_d+1})\cdots k_d(x_m,y_m) \tag{2.4}$$

is the reproducing kernel of $\mathcal{H}^{\otimes \mathbf{i}}$ where $\mathcal{H} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_d$.

Def. 10 is the natural generalization of the combinatorial definition of cumulants in $\mathbb{R}^d$ and Appendix 2.C.2 gives an equivalent definition via a generating function analogous to (2.2). However, our posthoc justification that these are the "right" definitions for cumulants in an RKHS are Theorems 2.3.1 and 2.3.2 that show that these kernelized cumulants have the same powerful properties as classic cumulants in $\mathbb{R}^d$ (Theorem 2.2.1).

Example 2.3.1 (Kernelized cumulants)**.** Let γ be a probability measure on $\mathcal{X}_1 \times \mathcal{X}_2$, with the RKHSs $(\mathcal{H}_1,k_1),(\mathcal{H}_2,k_2)$ given. Denote the random variables $K_1 = k_1(X_1,\cdot), K_2 = k_2(X_2,\cdot)$ where $(X_1,X_2) \sim \gamma$. Then the degree two kernelized cumulants are given as $\kappa^{(2,0)}_{k_1,k_2}(\gamma) = \mathbb{E}\left[K_1^{\otimes 2}\right] - \mathbb{E}\left[K_1\right]^{\otimes 2}$, $\kappa^{(1,1)}_{k_1,k_2}(\gamma) = \mathbb{E}\left[K_1 \otimes K_2\right] - \mathbb{E}\left[K_1\right] \otimes \mathbb{E}\left[K_2\right], \kappa^{(0,2)}_{k_1,k_2}(\gamma) = \mathbb{E}\left[K_2^{\otimes 2}\right] - \mathbb{E}\left[K_2\right]^{\otimes 2}$.

Inner Products of Cumulants. Computing inner products of moments is straightforward thanks to a nonlinear kernel trick, see Lemma 2.E.1 in the Appendix. For example, given two probability measures γ_1,γ_2 with corresponding random variables $(X_1,\ldots,X_d) \sim \gamma_1, (Y_1,\ldots,Y_d) \sim \gamma_2$ on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ and RKHSs $(\mathcal{H}_1,k_1),\ldots,(\mathcal{H}_d,k_d)$ on $\mathcal{X}_1,\ldots,\mathcal{X}_d$ with bounded kernels, and $\mathcal{H} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_d$, we can express:

$$\langle \mu^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma_1), \mu^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma_2)\rangle_{\mathcal{H}^{\otimes \mathbf{i}}} = \mathbb{E}_{\gamma_1\otimes\gamma_2} k_1(X_1,Y_1)^{i_1}\cdots k_d(X_d,Y_d)^{i_d}, \tag{2.5}$$

where $\mu^{\mathbf{i}}_{k_1,\cdots,k_d}$ is defined in Def. 7, and the expectation is taken over the product measure $\gamma_1 \otimes \gamma_2$.

Example 2.3.2. In the particular case of $d = 1$, (2.5) reduces to the well-known formula for the inner product of mean embeddings $\langle \mu^{(1)}_k(\gamma_1), \mu^{(1)}_k(\gamma_2)\rangle_{\mathcal{H}_k} = \mathbb{E}_{\gamma_1\otimes\gamma_2} k(X,Y)$.

Lemma 2.3.1 (Inner product of cumulants)**.** *Let $(\mathcal{H}_1,k_1),\ldots,(\mathcal{H}_d,k_d)$ be RKHSs with bounded kernels on $\mathcal{X}_1,\ldots,\mathcal{X}_d$ respectively, and let γ and η two probability measures on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_d$, $\mathbf{i} = (i_1,\ldots,i_d) \in \mathbb{N}^d$ such that $\deg(\mathbf{i}) = m$. Then*

$$\langle \kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma), \kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\eta)\rangle_{\mathcal{H}^{\otimes \mathbf{i}}} = \sum_{\pi,\tau\in P(m)} c_\pi c_\tau \mathbb{E}_{\gamma^{\mathbf{i}}_\pi \otimes \eta^{\mathbf{i}}_\tau} k^{\otimes \mathbf{i}}((X_1,\ldots,X_m),(Y_1,\ldots,Y_m)).$$

Point Separating Kernels. In the classic MMD setting the injectivity of the mean embedding $\gamma \mapsto \mathbb{E}_{X\sim\gamma}[k(X,\cdot)]$ on probability measures (known as the characteristic property of the kernel k) is equivalent to the MMD being a metric; this property is central in applications. We formulate our theoretical results in the next section using the much weaker property of what we term "point-separating" which is satisfied for essentially all popular kernels.

Definition 11 (Point-separating kernel)**.** We call a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ point-separating if the canonical feature map $\Phi : x \mapsto k(x,\cdot)$ is injective.

2.3.1 (Semi-)Metrics for Probability Measures

In this section we use cumulants to characterize probability measures and show how to compute the distance between kernelized cumulants with the expected kernel trick.

Theorem 2.3.1 (Characterization of distributions with cumulants)**.** *Let γ and η be two probability measures on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_d$, $(\mathcal{H}_1, k_1), \ldots, (\mathcal{H}_d, k_d)$ RKHSs on $\mathcal{X}_1, \ldots, \mathcal{X}_d$ such that for every $1 \le j \le d$ k_j is a bounded, continuous, point-separating kernel. Then*

$$\gamma = \eta \text{ if and only if } \kappa_{k_1,\ldots,k_d}(\gamma) = \kappa_{k_1,\ldots,k_d}(\eta).$$

Moreover, the expected kernel trick applies and for $\mathbf{i} \in \mathbb{N}^d$ with $\deg(\mathbf{i}) = m$, and $k^{\otimes\mathbf{i}}$ and $\mathcal{H}^{\otimes\mathbf{i}}$ as in (2.4)

$$\begin{aligned} d^{\mathbf{i}}(\gamma,\eta) &:= \|\kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma) - \kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\eta)\|^2_{\mathcal{H}^{\otimes\mathbf{i}}} \\ &= \sum_{\pi,\tau\in P(m)} c_\pi c_\tau \Big[\mathbb{E}_{\gamma^{\mathbf{i}}_\pi\otimes\gamma^{\mathbf{i}}_\tau} k^{\otimes\mathbf{i}}((X_1,\ldots,X_m),(Y_1,\ldots,Y_m)) \\ &\qquad + \mathbb{E}_{\eta^{\mathbf{i}}_\pi\otimes\eta^{\mathbf{i}}_\tau} k^{\otimes\mathbf{i}}((X_1,\ldots,X_m),(Y_1,\ldots,Y_m)) \\ &\qquad - 2\mathbb{E}_{\gamma^{\mathbf{i}}_\pi\otimes\eta^{\mathbf{i}}_\tau} k^{\otimes\mathbf{i}}((X_1,\ldots,X_m),(Y_1,\ldots,Y_m))\Big]. \end{aligned} \tag{2.6}$$

We recall Example 2.3.1 and now give examples of distances between such expressions.

Example 2.3.3 ($m = 1$)**.** Applied with $m = 1$ and $d = 1$, (2.6) becomes $\mathrm{MMD}^2_k(\gamma,\eta)$

$$\|\kappa^{(1)}_k(\gamma) - \kappa^{(1)}_k(\eta)\|^2_{\mathcal{H}_k} = \mathbb{E}k(X,X') + \mathbb{E}k(Y,Y') - 2\mathbb{E}k(X,Y),$$

where X, X' denotes independent copies of γ and Y, Y' denotes independent copies of η.

Example 2.3.4 ($m = 2$)**.** For $m = 2$ and $d = 1$, (2.6) reduces to

$$\begin{aligned}\|\kappa_k^{(2)}(\gamma) - \kappa_k^{(2)}(\eta)\|^2_{\mathcal{H}^{(1,1)}} = \mathbb{E}k(X,X')k(X'',X''') + \mathbb{E}k(Y,Y')k(Y'',Y''') + \mathbb{E}k(X,X')^2 \\ +\mathbb{E}k(Y,Y')^2 + 2\mathbb{E}k(X,Y)k(X',Y) + 2\mathbb{E}k(X,Y)k(X,Y') - 2\mathbb{E}k(X,Y)k(X',Y') \\ -2\mathbb{E}k(X,Y)^2 - 2\mathbb{E}k(X,X')k(X,X'') - 2\mathbb{E}k(Y,Y')k(Y,Y''),\end{aligned}$$

where X, X', X'', X''' denotes independent copies of γ and Y, Y', Y'', Y''' denotes independent copies of η. This expression compares the variances in the RKHS instead of the means. This is an example of the *kernel variance embedding* defined in the next subsection.

The price for the weak assumption of a point-separating kernel is that without any stronger assumptions one does not get a metric in general, and the all-purpose way to achieve a metric is to take an infinite sum over all $d^{\mathbf{i}}$'s. If we only use the degree $m = 1$ term $d^{\mathbf{i}}$ reduces to the well-known MMD formula which requires characteristicness to become a metric (see Example 2.3.3). There are two reasons why working under weaker assumptions is useful: firstly, if the underlying kernel is not characteristic this sum gives a structured way to incorporate finer information that discriminates the two distributions; an extreme case is the linear kernel $k(x, y) = \langle x, y\rangle$ which is point-separating, and in this case the sum reduces to the differences of classical cumulants. Secondly, under the stronger assumption of characteristicness one already has a metric after truncation at degree $m = 1$ (the classical MMD). However, in the finite-sample case adding higher degree terms can lead to increased power. Indeed, our experiments (Section 5.6) show that even just going one degree further (i.e. taking $m = 2$), can lead to more powerful tests.

2.3.2 A Characterization of Independence

Here we characterize independence in terms of kernelized cumulants.

Theorem 2.3.2 (Characterization of independence with cumulants)**.** *Let γ be a probability measure on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_d$, and $(\mathcal{H}_1, k_1), \ldots, (\mathcal{H}_d, k_d)$ RKHSs on $\mathcal{X}_1, \ldots, \mathcal{X}_d$ such that for every $1 \leq j \leq d$ k_j is a bounded, continuous, point-separating kernel. Then*

$$\gamma = \gamma|_{\mathcal{X}_1} \otimes \cdots \otimes \gamma|_{\mathcal{X}_d} \text{ if and only if } \kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma) = 0$$

for every $\mathbf{i} \in \mathbb{N}_+^d$. Moreover, the expected kernel trick applies in the sense that for $\mathbf{i} \in \mathbb{N}_+^d$

$$\|\kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma)\|^2_{\mathcal{H}^{\otimes \mathbf{i}}} = \sum_{\pi,\tau \in P(m)} c_\pi c_\tau \mathbb{E}_{\gamma^{\mathbf{i}}_\pi \otimes \gamma^{\mathbf{i}}_\tau} k^{\otimes \mathbf{i}}((X_1, \ldots, X_m), (Y_1, \ldots, Y_m)), \qquad (2.7)$$

where $m := \deg(\mathbf{i})$*, and* $k^{\otimes \mathbf{i}}$ *and* $\mathcal{H}^{\otimes \mathbf{i}}$ *are defined as in* (2.4)*.*

Applied to $\mathbf{i} = (1,1)$, the expression (2.7) reduces to the classical HSIC for two components, see Example 2.3.5 below. But for general $\mathbf{i}$ this construction leads to genuine new statistics in RKHSs.

Example 2.3.5 (Specific case: HSIC)**.** If $d = 2$£ there is only one order 2 index in $\mathbb{N}_+^d$, namely $\mathbf{i} = (1,1)$; in this case (2.7) reduces to the classical HSIC equation

$$\|\kappa_{k_1,k_2}^{(1,1)}(\gamma)\|_{\mathcal{H}^{(1,1)}}^2 = \mathbb{E}k_1(X,Y)k_2(X,Y) + \mathbb{E}k_1(X,Y)k_2(X',Y') - 2\mathbb{E}k_1(X,Y)k_2(X',Y),$$

where (X,Y) and (X',Y') are independent copies of the same random variable following γ.

2.3.3 Finite-Sample Statistics

To apply Theorem 2.3.1 and Theorem 2.3.2 in practice, one needs to estimate expressions such as $\mathbb{E}k^{\otimes \mathbf{i}}((X_1,\dots,X_m),(Y_1,\dots,Y_m))$. One could use classical estimators such as *U-statistic* [195] which lead to unbiased estimators. However, we follow [94] and use a *V-statistic* which is biased but conceptually simpler, easier, and efficient to compute. We note that the estimators presented here all have quadratic complexity like MMD and HSIC, see Appendix 2.E.

A two-sample test for non-characteristic feature maps. If k is characteristic then $\mathrm{MMD}_k(\gamma,\eta) = 0$ exactly when $\gamma = \eta$, but we can still increase testing power by considering the distance between the kernel variance and skewness embeddings, which leads us to use our semi-metrics $d^{(2)}(\gamma,\eta)$ and $d^{(3)}(\gamma,\eta)$ as defined in (2.6). An efficient estimator for $d^{(3)}$ is given in detail in Appendix 2.E; we provide the full expression for $d^{(2)}$ here.

Lemma 2.3.2 ($d^{(2)}$ estimation, see (2.6))**.** *The V-statistic for* $d^{(2)}(\gamma,\eta) = \|\kappa_k^{(2)}(\gamma) - \kappa_k^{(2)}(\eta)\|_{\mathcal{H}^{(1,1)}}^2$ *is*

$$\frac{1}{N^2}\mathrm{Tr}\Big[(\mathbf{K}_x\mathbf{J}_N)^2\Big] + \frac{1}{M^2}\mathrm{Tr}\Big[(\mathbf{K}_y\mathbf{J}_M)^2\Big] - \frac{2}{NM}\mathrm{Tr}\Big[\mathbf{K}_{xy}\mathbf{J}_M\mathbf{K}_{xy}^\top\mathbf{J}_N\Big],$$

where Tr *denotes trace,* $(x_n)_{n=1}^N \overset{\text{i.i.d.}}{\sim} \gamma$, $(y_m)_{m=1}^M \overset{\text{i.i.d.}}{\sim} \eta$, $\mathbf{K}_x = [k(x_i,x_j)]_{i,j=1}^N \in \mathbb{R}^{N\times N}$, $\mathbf{K}_y = [k(y_i,y_j)]_{i,j=1}^M \in \mathbb{R}^{M\times M}$, $\mathbf{K}_{x,y} = [k(x_i,y_j)]_{i,j=1}^{N,M} \in \mathbb{R}^{N\times M}$, $\mathbf{J}_n = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \in \mathbb{R}^{n\times n}$.

A kernel independence test. By Theorem 2.3.2, if $\gamma = \gamma|_{\mathcal{X}_1} \otimes \gamma|_{\mathcal{X}_2}$, then $\kappa^{(2,1)}(\gamma) = 0$ and $\kappa^{(1,2)}(\gamma) = 0$. We may compute the magnitude of either $\kappa^{(2,1)}(\gamma)$ or $\kappa^{(1,2)}(\gamma)$ – we will refer to these quantities as *cross skewness independence criterion* (CSIC). Note that these criteria are asymmetric. When $d = 2$ we have a probability measure γ on $\mathcal{X}_1 \times \mathcal{X}_2$ and two kernels $k : \mathcal{X}_1^2 \to \mathbb{R}$, $\ell : \mathcal{X}_2^2 \to \mathbb{R}$. Assume that we have samples $(x_i, y_i)_{i=1}^N$ and use the shorthand notation $\mathbf{K} = \mathbf{K}_x, \mathbf{L} = \mathbf{L}_y$ (similarly to Lemma 2.3.2) and $\mathbf{H} = \mathbf{H}_N$. Denote by $\circ$ the Hadamard product and $\langle \cdot \rangle$ the sum over all elements of a matrix. Then one can derive the following CSIC estimator.

Lemma 2.3.3 (CSIC estimation)**.** *The V-statistic for* $\|\kappa_{k,\ell}^{(1,2)}(\gamma)\|^2_{\mathcal{H}_k^{\otimes 1} \otimes \mathcal{H}_\ell^{\otimes 2}}$ *is*

$$\frac{1}{N^2}\Bigg\langle \mathbf{K} \circ \mathbf{K} \circ \mathbf{L} - 4\mathbf{K} \circ \mathbf{K}\mathbf{H} \circ \mathbf{L} - 2\mathbf{K} \circ \mathbf{K} \circ \mathbf{L}\mathbf{H} + 4\mathbf{K}\mathbf{H} \circ \mathbf{K} \circ \mathbf{L}\mathbf{H}$$
$$+ 2\mathbf{K} \circ \mathbf{L}\left\langle \frac{\mathbf{K}}{N^2} \right\rangle + 2\mathbf{K}\mathbf{H} \circ \mathbf{H}\mathbf{K} \circ \mathbf{L} + 4\mathbf{K} \circ \mathbf{H}\mathbf{K} \circ \mathbf{L}\mathbf{H} + \mathbf{K} \circ \mathbf{K}\left\langle \frac{\mathbf{L}}{N^2} \right\rangle$$
$$- 8\mathbf{K} \circ \mathbf{L}\mathbf{H}\left\langle \frac{\mathbf{K}}{N^2} \right\rangle - 4\mathbf{K} \circ \mathbf{H}\mathbf{K}\left\langle \frac{\mathbf{L}}{N^2} \right\rangle + 4\left\langle \frac{\mathbf{K}}{N^2} \right\rangle^2 \mathbf{L}\Bigg\rangle.$$

2.4 Experiments

In this section, we demonstrate the efficiency of the proposed kernel cumulants in two-sample and independence testing.[4]

- Two-sample test: Given $N - N$ samples from two probability measures γ and η on a space $\mathcal{X}$, the goal was to test the null hypothesis $H_0 : \gamma = \eta$ against the alternative $H_1 : \gamma \neq \eta$. The compared test statistics (S) were MMD, $d^{(2)}$, and $d^{(3)}$.
- Independence test: Given N paired samples from a probability measure γ on a product space $\mathcal{X}_1 \times \mathcal{X}_2$, the aim was to the test the null hypothesis $H_0 : \gamma = \gamma_1 \otimes \gamma_2$ against the alternative $H_1 : \gamma \neq \gamma_1 \otimes \gamma_2$. The compared test statistics (S) were HSIC and CSIC.

In our experiments H_1 held, and the estimated power of the tests is reported. Permutation test was applied to approximate the null distribution and its 0.95-quantile (which corresponds to the level choice $\alpha = 0.05$): We first computed our test statistic S using the given samples ($S_0 = S$), and then permuted the samples 100 times. If S_0 was in a high percentile ($\geq 95\%$ in our case) of the resulting distribution of S under the permutations, we rejected the null. We repeated these experiments

[4] All the code replicating our experiments is available at https://github.com/PatricBonnier/Kernelized-Cumulants.

100 times to estimate the power of the test. This procedure was in turn repeated 5 times and the 5 samples are plotted as a box plot along with a line plot showing the mean against the number of samples (N) used. All experiments were performed using the rbf-kernel

$$\mathrm{rbf}_\sigma(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2\sigma^2}},$$

where the parameter σ is called the *bandwidth*. We performed all experiments for every bandwidth of the form $\sigma = a10^b$ where $a = 1, 2.5, 5, 7.5$ and $b = -5, -4, -3, -2, -1, 0$ and the optimal value across the bandwidths was chosen for each method and sample size. The experiments were carried out on a laptop with an i7 CPU and 16GBs of RAM.

2.4.1 Synthetic Data

For synthetic data we designed two experiments.

- 2-sample test: We compared a uniform distribution with a mixture of two uniforms.
- Independence test: We considered the joint measure of a uniform and a correlated χ^2 random variable. We also use this same benchmark to compare the efficiency of classical and kernelized cumulants in Appendix 2.D.

Comparing a uniform with a mixture of uniforms. Even for simpler distributions like mixtures of uniform distributions it can be hard to pick up higher order features, and $d^{(2)}$ can outperform MMD even when provided with a moderate number of samples. Here we compared one uniform distribution $U[-1, 1]$ with an equal mixture of $U[0.35, 0.778]$ and $U[-0.35, -0.778]$. The endpoints in the mixture were chosen to match the first three moments of $U[-1, 1]$. The number of samples used ranged from 5 to 50, and the results are summarized in Fig. 2.1. One can see that with $d^{(2)}$ the power approaches 100% much faster than with using MMD.

Independence between a uniform and a χ^2. Let $X \sim U[0, 1]$ and $Z \sim N(0, 1)$ be independent of X. Denote by Φ the c.d.f. of a standard normal distribution and define Y_p to be a mixture with weight p at $\Phi^{-1}(X)$ and weight $1 - p$ at Z so that $Y_0 = Z$ and $Y_1 = \Phi^{-1}(X)$. We test for independence for $p = 0.5$ between Y_p^2—which will be χ^2 distributed with 1 degree of freedom–and X. As the statistical dependence of Y_p^2 and X is more complicated than a simple correlation we expect that higher order features of the data will help in the independence testing. The number of samples

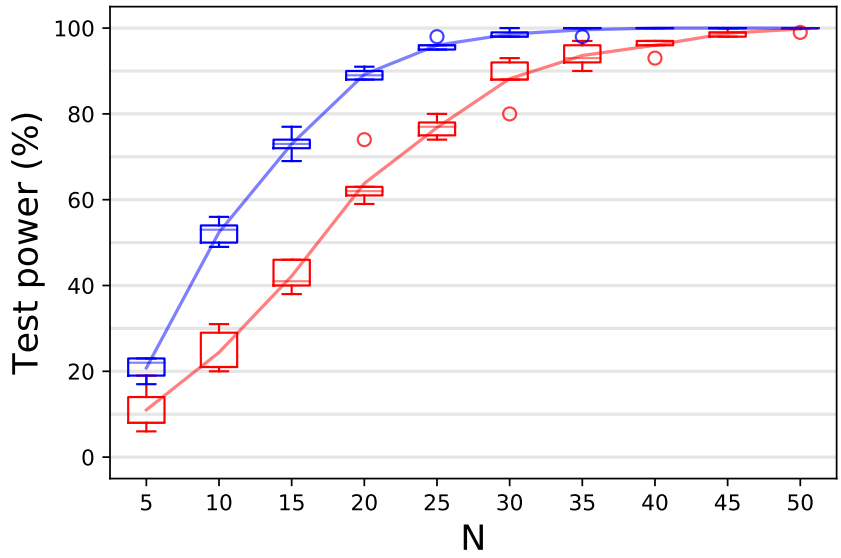


Figure 2.1: Test power as a function of the sample size (N) of two-sample test using the MMD (red) and $d^{(2)}$ statistics (blue), with $U[-1, 1]$ and an equal mixture of $U[0.35, 0.778]$ and $U[-0.35, -0.778]$ compared.

used ranged from 6 to 60, with the results summarized in Fig. 2.2. One can see that CSIC supersedes HSIC for every sample size, and the difference is more pronounced for smaller ones.

2.4.2 Real-World Data

We demonstrate the efficiency of the kernelized cumulants on real-world data. We designed two experiments.

- 2-sample test: Here the goal was to test if environmental data in two different seasons, and traffic data at different speeds can be distinguished.
- independence test: The aim was to test if two distributions describing traffic flow and other traffic factors are independent.

To improve performance of both test statistics, all features are standardized to lie between 0 and 1.

Seoul bicycle data. The Seoul bicycle data set [63] consists of environmental data along with the number of bicycle rentals. The environmental data consists of 9 numerical values, 1 categorical value (season), and two binary values. We compare the distribution of the environmental data in the winter and the fall, as we expect these distributions to be different. Concretely, we do 2-sample testing on two measures γ, η

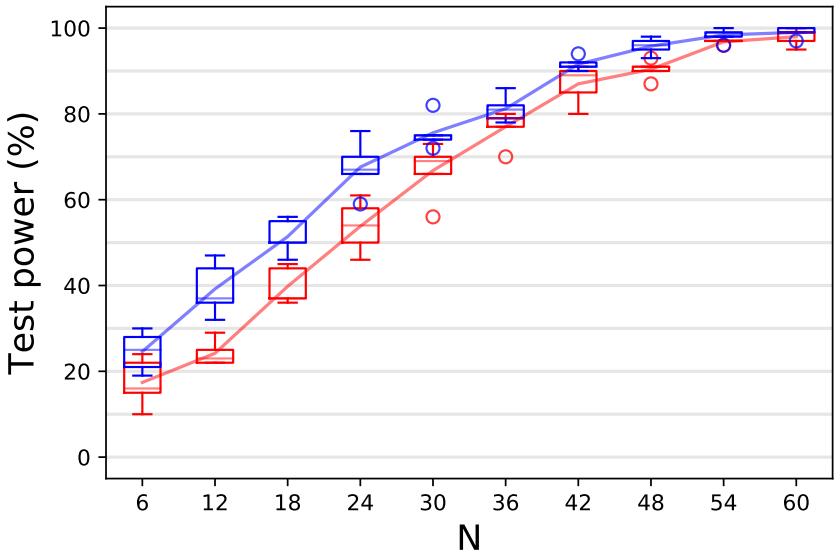


Figure 2.2: Test power as a function of the sample size (N) of independence testing using HSIC (red) and CSIC statistics (blue), with the independence testing between $Y^2_{0.5}$ and X.

on $\mathbb{R}^{11}$, and assume that $\gamma \neq \eta$ where γ is the distribution of the environmental data in winter, and η is that of the data in the fall. Permutation testing was performed for N between 4 and 44, with results summarized in Fig. 2.3. As it can be observed that $d^{(2)}$ outperforms MMD in terms of test power.

For the type I error, i.e. the probability of falsely rejecting the null hypothesis (comparing winter data with itself), it hovers between $5-10\%$ for both statistics, which is admittedly slightly higher than the desired 5% due to the small sample size, but very similar for both statistics. The results for the Type I error are summarized in Fig. 2.8 in Appendix 2.D.

Brazilian traffic data. We used the Sao Paulo traffic benchmark [68] to perform independence testing. The dataset consists of 16 different integer-valued statistics about the hourly traffic in Sao Paulo such as blockages, fires and other reasons that might hold up traffic. This is combined with a number that describes the slowness of traffic at the given hour; so $\mathcal{X}_1 = \mathbb{R}^{16}$, $\mathcal{X}_2 = \mathbb{R}$. One expects a strong dependence between the two sets—or equivalently, for the null hypothesis to be false—and for the statistics are heavily skewed towards 0 as it is naturally sparse. For independence testing we performed permutation testing for N between 4 and 40. The resulting test powers are summarized in Fig. 2.4. As it can be seen, HSIC and CSIC performs

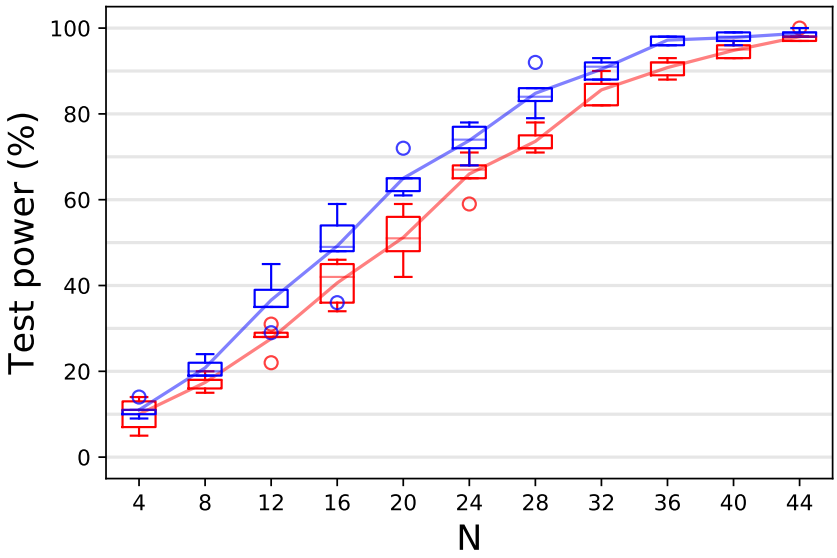


Figure 2.3: Two-sample testing using MMD (red) and $d^{(2)}$ (blue) on the Seoul bicycle data set.

similarly for very low sample sizes, but for anything else CSIC is the favorable statistic in terms of test power.

For two-sample testing, we sampled N between 5 and 50 and compared the distribution of slow moving traffic with the fast moving traffic. The results are summarized in Fig. 2.5. It is clear that $d^{(3)}$ performs similarly to MMD in terms of test power for very small sample sizes, but significantly better for larger ones.

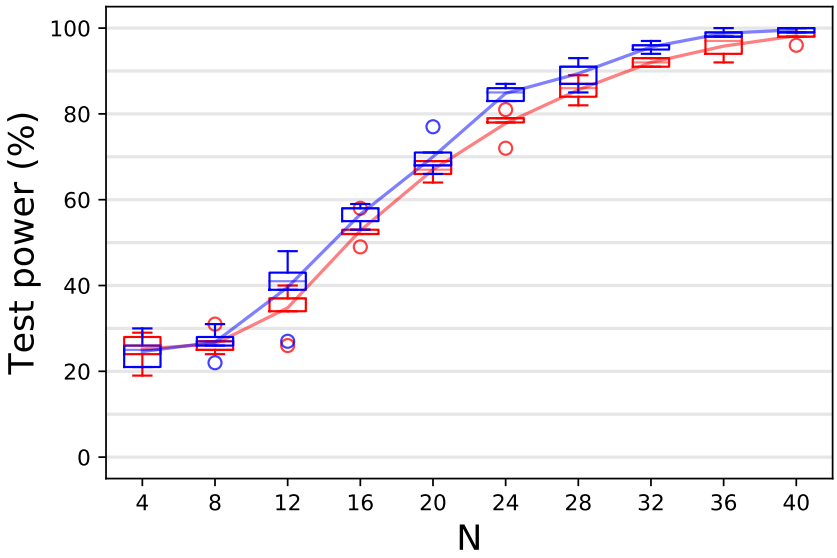


Figure 2.4: Independence testing using HSIC (red) and CSIC (blue) on the Sao Paulo traffic dataset.

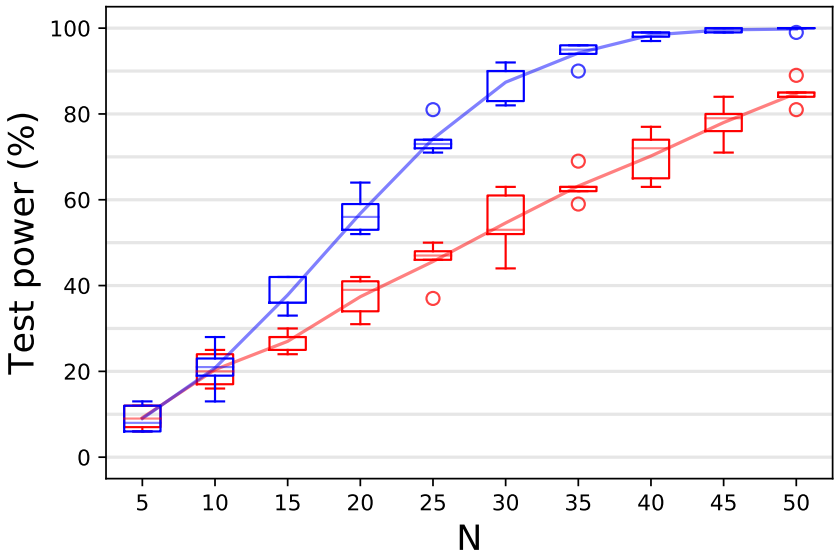


Figure 2.5: Two-sample testing using MMD (red) and $d^{(3)}$ (blue) on the Sao Paulo traffic dataset.

2.5 Conclusion

We defined cumulants for random variables in RKHSs by extending the algebraic characterization of cumulants on $\mathbb{R}^d$. This results in a structured description of the law of random variables that goes beyond the classic kernelized mean and covariance.

A kernel trick allows us to compute these kernelized cumulants. We applied our theoretical results to two-sample and independence testing; although kernelized mean and covariance are sufficient for this task, the higher-order kernelized cumulants have the potential to increase the test power and to relax the assumptions on the kernel. Our experiments on real and synthetic data show that this can indeed lead to large increases in test power. A disadvantage of these higher order statistics is that their theoretical analysis requires more mathematical machinery although we emphasize that the resulting estimators are simple V-statistics.

These appendices provide additional background and elaborate on some of the finer points in the main text. In Appendix 2.A we illustrate that cumulants have typically lower variance estimators compared to moments. Technical background on tensor products and tensor sums of Hilbert spaces, and on tensor algebras is provided in Appendix 2.B. We present our proofs in Appendix 2.C. In Appendix 2.D additional details on our numerical experiments are provided. Our V-statistic based estimators are detailed in Appendix 2.E.

2.A Moments and Cumulants

Moments have well-known drawbacks. For example, one of their undesirable properties is that the moments of lower degree can dominate moments of higher degree which is in particular troublesome if finite samples are used to estimate the moment sequence. Even in dimension $d = 1$ the second moment is dominated by the squared mean, that is for a real-valued random variable $X \sim \gamma$

$$\mu^2(\gamma) = (\mu^1(\gamma))^2 + \mathrm{Var}(X),$$

where $\mathrm{Var}(X) := \mathbb{E}[(X - \mu^1(\gamma))^2]$. It is well known that the minimum variance unbiased estimators for the variance are more efficient than that for the second moment: denoting them by $\widehat{\mu^2}$ and $\widehat{\kappa}$ respectively, one can show [30] that given N samples from X, the following holds

$$\mathrm{Var}\left(\widehat{\mu^2}\right) = \mathrm{Var}\left(\widehat{\kappa}\right) + \frac{2}{N}\left[(\mathbb{E}X)^4 - (\mathbb{E}X)^2\,\mathrm{Var}(X) - 2\frac{\mathrm{Var}(X)^2}{N-1}\right]$$

so that when X has a large mean, it is more efficient to estimate its variance than its second moment since the last term in the above expression dominates. Hence, the variance $\mathrm{Var}(X)$ is typically a much more sensible second-order statistic than $\mu^2(\gamma)$.

Further, we note that although cumulants are classic for vector-valued data, there seems to be not much work done about extending their properties to general structured data. Our kernelized cumulants apply to any set $\mathcal{X}$ for which a positive definite kernel is given; this includes many practically relevant examples such as strings [127], graphs [116], or general sequentially ordered data [113, 46]; see example [81] for survey of kernels for structured data.

2.B Technical Background

In Section 2.B.1 the tensor products $(\bigotimes_{j=1}^{d} \mathcal{H}_j)$ and direct sums of Hilbert spaces $(\bigoplus_{i\in I} \mathcal{H}_i)$ are recalled. Section 2.B.2 is about tensor algebras over Hilbert spaces $(\prod_{m\geq 0} \mathcal{H}^{\otimes m})$.

2.B.1 Tensor Products and Direct Sums of Banach and Hilbert Spaces

Tensor products of Hilbert spaces. For Hilbert spaces $\mathcal{H}, \ldots, \mathcal{H}_d$ and $(h_1, \ldots, h_d) \in \mathcal{H}_1 \times \cdots \times \mathcal{H}_d$, the multi-linear operator $h_1 \otimes \cdots \otimes h_d \in \mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_d$ is defined as

$$(h_1 \otimes \cdots \otimes h_d)(f_1, \ldots, f_d) = \prod_{j=1}^{d} \langle h_j, f_j \rangle_{\mathcal{H}_j}$$

for all $(f_1, \ldots, f_d) \in \mathcal{H}_1 \times \cdots \times \mathcal{H}_d$. By extending the inner product

$$\langle a_1 \otimes \cdots \otimes a_d, b_1 \otimes \cdots \otimes b_d \rangle_{\mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_d} := \prod_{j=1}^{d} \langle a_j, b_j \rangle_{\mathcal{H}_j}$$

to finite linear combinations of $a_1 \otimes \cdots \otimes a_d$-s

$$\left\{ \sum_{i=1}^{n} c_i \otimes_{j=1}^{d} a_{i,j} \ : \ c_i \in \mathbb{R}, a_{i,j} \in \mathcal{H}_j, \ n \geq 1 \right\}$$

by linearity, and taking the topological completion one arrives at $\mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_d$. Specifically, if $(\mathcal{H}_1, k_1), \ldots, (\mathcal{H}_d, k_d)$ are RKHSs, then so is $\mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_d = \mathcal{H}_{\otimes_{j=1}^{d} k_j}$ [17, Theorem 13] with the tensor product kernel

$$\left(\otimes_{j=1}^{d} k_j \right) \left((x_1, \ldots, x_d), (x'_1, \ldots, x'_d) \right) := \prod_{j=1}^{d} k_j \left(x_j, x'_j \right)$$

for $(x_1, \ldots, x_d), (x'_1, \ldots, x'_d) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$.

Tensor products of Banach spaces. For Banach spaces $\mathcal{B}_1, \ldots \mathcal{B}_d$, the construction of $\mathcal{B}_1 \otimes \cdots \otimes \mathcal{B}_d$ is a little more involved [117] as one cannot rely on an inner product.

Direct sums of Hilbert and Banach spaces. Let $(\mathcal{H}_i)_{i\in I}$ be Hilbert or Banach spaces where I is some index set. The direct sum of $\mathcal{H}_i$-s— written as $\bigoplus_{i\in I} \mathcal{H}_i$— consists of ordered tuples $h = (h_i)_{i\in I}$ such that $h_i \in \mathcal{H}_i$ for all $i \in I$ and $h_i = 0$ for all but a finite number of $i \in I$. Operations (addition, scalar multiplication) are performed coordinate-wise, and the inner product of $a, b \in \bigoplus_{i\in I} \mathcal{H}_i$ is defined as $\langle a, b \rangle_{\bigoplus_{i\in I} \mathcal{H}_i} = \sum_{i\in I} a_i b_i$.

2.B.2 Tensor Algebras

The tensor algebra T_j over a Hilbert space $\mathcal{H}_j$ is defined as the topological completion of the space

$$\bigoplus_{m\geq 0} \mathcal{H}_j^{\otimes m}.$$

Note that it can equivalently be defined as the subset of $(h_0, h_1, h_2, \ldots) \in \prod_{m\geq 0} \mathcal{H}_j^{\otimes m}$ such that $\sum_{m\geq 0} \|h_m\|^2_{\mathcal{H}_j^{\otimes m}} < \infty$, and as such it is a Hilbert space with norm

$$\|(h_0, h_1, h_2, \ldots)\|^2_{\prod_{m\geq 0} \mathcal{H}_j^{\otimes m}} = \sum_{m\geq 0} \|h_m\|^2_{\mathcal{H}_j^{\otimes m}}.$$

T_j is also an algebra, endowed with the tensor product over $\mathcal{H}_j$ as its product. For $a = (a_0, a_1, a_2, a_2 \ldots), b = (b_0, b_1, b_2, b_2 \ldots) \in \mathrm{T}_j$, their product can be written down in coordinates as

$$a \cdot b = \left(\sum_{i=0}^{m} a_i \otimes b_{m-i} \right)_{m\geq 0}.$$

For a sequence $\mathcal{H}_1, \ldots, \mathcal{H}_d$ of Hilbert spaces, we define

$$\mathrm{T} := \mathrm{T}_1 \otimes \cdots \otimes \mathrm{T}_d,$$

where $\mathrm{T}_j = \prod_{m\geq 0} \mathcal{H}_j^{\otimes m}$ $(j = 1, \ldots, d)$. Let $\mathcal{H} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_d$, and recall that given a tuple of integers $\mathbf{i} = (i_1, \ldots, i_d) \in \mathbb{N}^d$ we define $\mathcal{H}^{\otimes \mathbf{i}} := \mathcal{H}_1^{\otimes i_1} \otimes \cdots \otimes \mathcal{H}_d^{\otimes i_d}$. This allows us to write down a multi-grading for T as

$$\mathrm{T} = \prod_{\mathbf{i}\in\mathbb{N}^d} \mathcal{H}^{\otimes \mathbf{i}}.$$

Note that this gives credence to us using multi-indices $\mathbf{i} \in \mathbb{N}^d$ to describe elements of the tensor algebra, as the multi-indices form its multi-grading.

Furthermore, T is a multi-graded algebra when endowed with the (linear extension of the) following multiplication defined on the components of T

$$\begin{aligned} &\star : \mathcal{H}^{\otimes \mathbf{i}^1} \times \mathcal{H}^{\otimes \mathbf{i}^2} \to \mathcal{H}^{\otimes (\mathbf{i}^1 + \mathbf{i}^2)}, \\ &(x_1 \otimes \cdots \otimes x_d) \star (y_1 \otimes \cdots \otimes y_d) = (x_1 \cdot y_1) \otimes \cdots \otimes (x_d \cdot y_d), \end{aligned} \tag{2.8}$$

so that for $a = \left(a^{\mathbf{i}}\right)_{\mathbf{i}\in\mathbb{N}^d}, b = \left(b^{\mathbf{i}}\right)_{\mathbf{i}\in\mathbb{N}^d} \in \mathrm{T}$, their product can be written down as

$$(a \star b)^{\mathbf{i}} = \sum_{\mathbf{i}^1 + \mathbf{i}^2 = \mathbf{i}} a^{\mathbf{i}^1} \star b^{\mathbf{i}^2} \tag{2.9}$$

where addition of tuples $\mathbf{i}^1, \mathbf{i}^2 \in \mathbb{N}^d$ is defined as $\mathbf{i}^1 + \mathbf{i}^2 = (i_1^1 + i_1^2, \ldots, i_d^1 + i_d^2)$. With the degree of a tuple defined as $\deg(\mathbf{i}) = i_1 + \cdots + i_d$, T is also a graded algebra, with the grading written down as

$$\mathrm{T} = \prod_{m \geq 0} \bigoplus_{\{\mathbf{i} \in \mathbb{N}^d : \deg(\mathbf{i}) = m\}} \mathcal{H}^{\otimes \mathbf{i}},$$

so that if you multiply two elements together, the degree of their product is the sum of their degree. Finally we note that T is a unital algebra and the unit has the explicit form

$$(1, 0, 0, \ldots),$$

i.e. the element consisting of only a 1 at degree 0.

2.C Proofs

This section is dedicated to proofs. The equivalence between the combinatorial expressions of cumulants and the definition via a moment generating function is proved in Section 2.C.2. The derivation of our main results (Theorem 2.3.1 and Theorem 2.3.2) are detailed in Section 2.C.3.

2.C.1 Equivalent Definitions of Cumulants in $\mathbb{R}^d$

Here we introduce a classical definition of cumulants via a moment generating function and its equivalence to the combinatorial expressions. If $X = (X_1, \ldots, X_d)$ is an $\mathbb{R}^d$-valued random variable distributed according to $X \sim \gamma$, then

$$\mu^{\mathbf{i}} = \mathbb{E}[X_1^{i_1} \cdots X_d^{i_d}] \in \mathbb{R}$$

for $\mathbf{i} = (i_1, \ldots, i_d) \in \mathbb{N}^d$. The following definition of the cumulants $\kappa^{\mathbf{i}}(\gamma)$ of γ are equivalent

1. $\sum_{\mathbf{i} \in \mathbb{N}^d} \kappa^{\mathbf{i}}(\gamma) \frac{\theta^{\mathbf{i}}}{\mathbf{i}!} = \log \sum_{\mathbf{i} \in \mathbb{N}^d} \mu^{\mathbf{i}}(\gamma) \frac{\theta^{\mathbf{i}}}{\mathbf{i}!}$,
2. $\kappa^{\mathbf{i}}(\gamma) = \sum_{\pi \in P(d)} c_\pi \prod_{\sigma \in \pi} \mu^{\mathbf{i}}(\sigma)$,

where $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$, $c_\pi = (-1)^{|\pi|}(|\pi| - 1)!$ and the product $\prod_{\sigma \in \pi}$ is over all the blocks $\sigma \in (\pi_1, \ldots, \pi_b)$ in the partition $\pi = (\pi_1, \ldots, \pi_b)$ of $\{1, \ldots, d\}$. The equivalence between these two definitions of cumulants, via a generating function and via their combinatorial definition, is classic [133]. This equivalence is also at the heart of many proofs about properties of cumulants since some properties are easier to prove via one or the other definition.

2.C.2 Equivalent Definitions of Cumulants in RKHS

In the main text, we defined cumulants in RKHS by mimicking the combinatorial definition of cumulants in $\mathbb{R}^d$. It is natural and useful to also have the analogous definition via a "generating function" for RKHS-valued random variables. However, to generalize the definition via the logarithm of the moment generating function to random variables in RKHS, requires to define a logarithm for tensor series of moments. In this part, we show that this can be done and that indeed the two definitions are equivalent.

We use the shorthand $\kappa(\gamma) := \kappa_{k_1,\dots,k_d}(\gamma)$, $\mu(\gamma) := \mu_{k_1,\dots,k_d}(\gamma)$, and we overload the notation $(X_1, \dots, X_d)$ with $(k_1(\cdot, X_1), \dots, k_d(\cdot, X_d))$. With this notation, we show that given coordinates $\mathbf{i} \in \mathbb{N}^d$, one may express the generalized cumulant $\kappa^{\mathbf{i}}(\gamma)$ as either a combinatorial sum over moments indexed by partitions, or by using the cumulant generating function.

More specifically, we show that the generalized cumulant of a probability measure γ on $\mathcal{H}_1 \times \cdots \times \mathcal{H}_d$ defined as

$$\kappa^{\mathbf{i}}(\gamma) = \sum_{\pi \in P(m)} c_\pi \mathbb{E}_{\gamma_\pi^{\mathbf{i}}}(X^{\otimes \mathbf{i}})$$

where $c_\pi = (-1)^{|\pi|-1}(|\pi| - 1)!$ can also be expressed as coordinates in the tensorized logarithm of the moment series. Motivated by the Taylor series expansion of the classic logarithm, we define

$$\log : \mathrm{T} \to \mathrm{T}, \quad x \mapsto \sum_{n \geq 1} \frac{(-1)^{n-1}}{n} (x - 1)^{\star n},$$

where $\star$ denotes the product as defined in (2.8) and for $t \in \mathrm{T}$, $t^{\star n}$ is defined as

$$t^{\star n} = \underbrace{t \star \cdots \star t}_{\text{n - times}},$$

or coordinate-wise $(t^{\star n})^{\mathbf{i}} = \sum_{\mathbf{i}^1 + \cdots + \mathbf{i}^n = \mathbf{i}} t^{\mathbf{i}^1} \star \cdots \star t^{\mathbf{i}^n}$ for $\mathbf{i} \in \mathbb{N}^d$. Note that unlike the classical logarithm $\log : \mathbb{R}_+ \to \mathbb{R}$, the tensorized logarithm is defined on the whole space as a formal expression.

Generalized Cumulants as Logarithms We want to show that the following holds

$$\kappa^{\mathbf{i}}(\gamma) = \Big(\log \mu(\gamma^{\mathbf{i}}) \Big)^{\mathbb{1}_m}, \tag{2.10}$$

where $\mathbb{1}_m = (1, \dots, 1) \in \mathbb{N}^m$. By iterating (2.9) we can express (2.10) as

$$\sum_{j=1}^{m} \frac{(-1)^{j-1}}{j} \sum_{\mathbf{i}^1 + \cdots + \mathbf{i}^j = \mathbb{1}_m} \mu^{\mathbf{i}^1}(\gamma^{\mathbf{i}}) \star \cdots \star \mu^{\mathbf{i}^j}(\gamma^{\mathbf{i}}),$$

and our goal is to express this as a sum over partitions. We will use the notation $[n] = \{1, \dots, n\}$. We can achieve our goal in two parts:

1. Show that for a fixed $\mathbf{i} \in \mathbb{N}^d$ with $\deg(\mathbf{i}) = m$ we can express (2.10) as a sum over all surjective functions from $[m]$ to $[j]$.

2. Show that this sum over functions reduces to a sum over partitions.

Part 1. Note that given $\mathbf{i}^1 + \cdots + \mathbf{i}^j = \mathbb{1}_m$ we may define $h : [m] \to [j]$ by the relation $(\mathbf{i}^{h(n)})_n = 1$, that is, we take $h(n)$ to be the index c for which the multi-index $\mathbf{i}^c$ is 1 at n. Note that this function is necessarily surjective since the sum is taken over non-zero multi-indices. Equivalently, for any surjective function $h : [m] \to [j]$ we may define multi-indices by setting

$$(\mathbf{i}^c)_n = \begin{cases} 1 & \text{if } n \in h^{-1}(c) \\ 0 & \text{otherwise} \end{cases}.$$

Note that any such multi-index will be non-zero since the function is assumed to be surjective. With this identification we can write

$$\Big(\log \mu(\gamma^{\mathbf{i}})\Big)^{\mathbb{1}_m} = \sum_{j=1}^{m} \frac{(-1)^{j-1}}{j} \sum_{h:[m]\to[j]} \mu^{\mathbf{i}^{h^{-1}(1)}}(\gamma^{\mathbf{i}}) \star \cdots \star \mu^{\mathbf{i}^{h^{-1}(j)}}(\gamma^{\mathbf{i}}).$$

Part 2. Recall that given a function $h : [m] \to [j]$ we can associate it to its corresponding partition $\pi_h \in \mathcal{P}(m)$ by considering the set $\{h^{-1}(1), \dots, h^{-1}(j)\}$, and there are exactly $j!$ different functions corresponding to a given partition, which are given by re-ordering the values $1, \dots, j$. This reordering of the blocks does not change the summands since the marginals of the partition measure are always copies of each other and hence self-commute, hence a product of moments like $\mu^{\mathbf{i}^{h^{-1}(1)}}(\gamma^{\mathbf{i}}) \star \cdots \star \mu^{\mathbf{i}^{h^{-1}(j)}}(\gamma^{\mathbf{i}})$ can always be written as $\mu^{\mathbf{i}}(\gamma^{\mathbf{i}}_{\pi_h})$, the $\mathbf{i}$-th coordinate of the moment sequence of the partition measure $\gamma^{\mathbf{i}}_{\pi_h}$. With this in mind we can write

$$\begin{aligned}\Big(\log \mu(\gamma^{\mathbf{i}})\Big)^{\mathbb{1}_m} &= \sum_{j=1}^{m} \frac{(-1)^{j-1}}{j} \sum_{h:[m]\to[j]} \mu^{\mathbf{i}}(\gamma^{\mathbf{i}}_{\pi_h}) = \sum_{\pi \in P(m)} \frac{(-1)^{|\pi|-1}}{|\pi|} |\pi|! \mu^{\mathbf{i}}(\gamma^{\mathbf{i}}_{\pi}) \\ &= \sum_{\pi \in P(m)} c_\pi \mu^{\mathbf{i}}(\gamma^{\mathbf{i}}_{\pi}) = \sum_{\pi \in P(m)} c_\pi \mathbb{E}_{\gamma^{\mathbf{i}}_{\pi}}(X^{\otimes \mathbf{i}}).\end{aligned}$$

From this it immediately follows that for two probability measures γ, η we can write

$$\begin{aligned}\langle \kappa^{\mathbf{i}}(\gamma), \kappa^{\mathbf{i}}(\eta)\rangle_{\mathcal{H}^{\otimes \mathbf{i}}} &= \Big\langle \sum_{\pi \in P(m)} c_\pi \mathbb{E}_{\gamma_\pi^{\mathbf{i}}}(X^{\otimes \mathbf{i}}), \sum_{\tau \in P(m)} c_\tau \mathbb{E}_{\eta_\tau^{\mathbf{i}}}(Y^{\otimes \mathbf{i}})\Big\rangle_{\mathcal{H}^{\otimes \mathbf{i}}} \\ &= \sum_{\pi,\tau \in P(m)} c_\pi c_\tau \mathbb{E}_{(X,Y)\sim \gamma_\pi^{\mathbf{i}} \otimes \eta_\tau^{\mathbf{i}}} \langle X^{\otimes \mathbf{i}}, Y^{\otimes \mathbf{i}}\rangle_{\mathcal{H}^{\otimes \mathbf{i}}}.\end{aligned}$$

Lemma 2.3.1 then follows from the definition of the tensor products.

2.C.3 Proof of Theorem 2.3.1 and Theorem 2.3.2

In this section we present the proofs of Theorem 2.3.1 and Theorem 2.3.2. We do this in a slightly more abstract setting where the feature maps take values in Banach spaces for clarity, until the end when we again restrict our attention to RKHSs. We start out by showing that polynomial functions of the feature maps characterize measures (Lemma 2.C.1). From there it is straightforward to show that cumulants have the same property (Theorem 2.C.1), and lastly that this also holds when working directly with the kernels (Proposition 2.C.2).

A *monomial* on separable Banach spaces $\mathcal{B}_1, \ldots, \mathcal{B}_d$ is any expression of the form

$$M(x_1, \ldots, x_d) = \prod_{j=1}^{i_1} \langle f_j^1, x_1\rangle \cdots \prod_{j=1}^{i_d} \langle f_j^d, x_d\rangle$$

for some $(i_1, \ldots, i_d) \in \mathbb{N}^d$, where $f_j^i \in \mathcal{B}_i^\star$ are elements of the dual space $\mathcal{B}_i^\star$ and $x_i \in \mathcal{B}_i$.[5] Finite linear combinations of monomials are called the *polynomials*. Recall that a set of functions F on a set S is said to *separate the points* of S if for every $x \neq y \in S$ there exists $f \in F$ such that $f(x) \neq f(y)$.

Lemma 2.C.1 (Polynomial functions of feature maps characterize probability measures)**.** *Let $\mathcal{X}_1, \ldots, \mathcal{X}_d$ be Polish spaces, $\mathcal{B}_1 \ldots, \mathcal{B}_d$ separable Banach spaces and $\varphi_i : \mathcal{X}_i \to \mathcal{B}_i$ be continuous, bounded, and injective functions. Then the set of functions on the Borel probability measures $\mathcal{P}\left(\prod_{i=1}^d \mathcal{X}_i\right)$ of $\prod_{i=1}^d \mathcal{X}_i$*

$$\mathcal{P}\left(\prod_{i=1}^d \mathcal{X}_i\right) \to \mathbb{R}, \quad \gamma \mapsto \int_{\prod_{i=1}^d \mathcal{X}_i} p\big(\varphi_1(x_1), \ldots, \varphi_d(x_d)\big) \mathrm{d}\gamma(x_1, \ldots, x_d),$$

where p ranges over all polynomials, separates the points of $\mathcal{P}\left(\prod_{i=1}^d \mathcal{X}_i\right)$.

Proof. We first show that the pushforward map

$$\prod_{i=1}^d \varphi_i : \mathcal{P}\left(\prod_{i=1}^d \mathcal{X}_i\right) \to \mathcal{P}\left(\prod_{i=1}^d \mathcal{B}_i\right)$$

[5] These monomials naturally extend the classical ones.

is injective. This is done in two parts, first we show that every Borel measure on $\prod_{i=1}^d \mathcal{X}_i$ is a Radon measure, then we show that the pushforward map is injective on Radon measures. To see the first part, note that since $\mathcal{X}_1, \dots, \mathcal{X}_d$ are Polish spaces, so is their product space $\prod_{i=1}^d \mathcal{X}_i$ ([61, Theorem 2.5.7]; [202, Theorem 16.4c]), and since Borel measures on Polish spaces are Radon measures [24, Theorem 7.1.7], any $\gamma \in \mathcal{P}(\prod_{i=1}^d \mathcal{X}_i)$ must be a Radon measure.

For the second part, note that

$$\prod_{i=1}^d \varphi_i : \prod_{i=1}^d \mathcal{X}_i \to \prod_{i=1}^d \mathcal{B}_i, \quad \left(\prod_{i=1}^d \varphi_i\right)(x_1, \dots, x_d) \mapsto \prod_{i=1}^d \varphi_i(x_i)$$

is a norm bounded, continuous injection. Since $\prod_{i=1}^d \mathcal{B}_i$ is a Hausdorff space, $\prod_{i=1}^d \varphi_i$ is a homeomorphism on compacts since continuous injections into Hausdorff spaces are homeomorphisms on compacts [153, Theorem 4.17]. Let $\mu, \nu \in \mathcal{P}\left(\prod_{i=1}^d \mathcal{X}_i\right)$ be two Radon measures such that their pushforwards are the same $\prod_{i=1}^d \varphi_i(\mu) = \prod_{i=1}^d \varphi_i(\nu)$, then for any compact $C \subseteq \prod_{i=1}^d \mathcal{X}_i$ we have $\mu(C) = \nu(C)$ as $\prod_{i=1}^d \varphi_i : C \to \prod_{i=1}^d \varphi_i(C)$ is a homeomorphism. Since Radon measures are characterized by their values on compacts, this implies that $\mu = \nu$. Hence the pushforward map is injective.

Denote by K the image of $\prod_{i=1}^d \mathcal{X}_i$ under the mapping $\prod_{i=1}^d \varphi_i$ in $\prod_{i=1}^d \mathcal{B}_i$. Note that K is a bounded Polish space. It is enough to show that the polynomials separate the points of $\mathcal{P}(K)$. To see this, note that the polynomials form an algebra of continuous functions that separate the points of $\prod_{i=1}^d \mathcal{B}_i$, and when restricted to K they are bounded, since K is norm bounded. Since K is Polish, any Borel measure is Radon, and we can apply the Stone-Weierstrass theorem for Radon measures [24, Exercise 7.14.79] to get the assertion. □

In what follows we will use the following index notation for linear functionals. Fix some tuple $\mathbf{i} = (i_1, \dots, i_d) \in \mathbb{N}^d$ with $\deg(\mathbf{i}) = m$. Given separable Banach spaces $\mathcal{B}_1 \dots, \mathcal{B}_d$ we use the notation

$$\mathcal{B}^{\otimes \mathbf{i}} := \mathcal{B}_1^{\otimes i_1} \otimes \cdots \otimes \mathcal{B}_d^{\otimes i_d}$$

and given an element $x = (x_1, \dots, x_d) \in \prod_{i=1}^d \mathcal{B}_i$ we write $x^{\mathbf{i}} := x_1^{\otimes i_1} \otimes \cdots \otimes x_d^{\otimes i_d}$ so that $x^{\mathbf{i}} \in \mathcal{B}^{\otimes \mathbf{i}}$. If we have functions $(\varphi_i)_{i=1}^d$ such that $\varphi_i : \mathcal{X}_i \to \mathcal{B}_i$ on some Polish spaces $\mathcal{X}_1, \dots, \mathcal{X}_d$, then we write

$$\varphi^{\otimes \mathbf{i}} := \varphi_1^{\otimes i_1} \otimes \cdots \otimes \varphi_d^{\otimes i_d}, \quad \varphi^{\otimes \mathbf{i}} : \prod_{i=1}^d \mathcal{X}_i \to \mathcal{B}^{\otimes \mathbf{i}}.$$

Given a collection of linear functionals $F \in \prod_{j=1}^{d} \left(\mathcal{B}_j^\star\right)^{i_j}$ such that $F = (f_1, \ldots, f_d)$ we write

$$F^{\otimes \mathbf{i}} := f_1 \otimes \cdots \otimes f_d, \quad F^{\otimes \mathbf{i}} \in \left(\mathcal{B}^{\otimes \mathbf{i}}\right)^\star.$$

Note the following trick: the monomials on $\prod_{i=1}^{d} \mathcal{B}_i$ are exactly functions of the form

$$x \mapsto \langle F^{\otimes \mathbf{i}}, x^{\mathbf{i}} \rangle$$

for $F = (f_1, \ldots, f_d)$, this will be used in the proofs. We can now restate and prove the our theorem. Note that the cumulants here are defined like in Definition 10 which is a sensible definition even if the feature maps are not associated to kernels.

Theorem 2.C.1 (Generalization of Theorem 2.3.1 and Theorem 2.3.2)**.** *Let $\mathcal{X}_1, \ldots, \mathcal{X}_d$ be Polish spaces and $\varphi_i : \mathcal{X}_i \to \mathcal{B}_i$ be continuous, bounded and injective feature maps into separable Banach spaces $\mathcal{B}_i$ for $i = 1, \ldots d$. Let γ and η be probability measures on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_d$. Then*

1. *$\gamma = \eta$ if and only if $\kappa(\gamma) = \kappa(\eta)$.*
2. *$\gamma = \bigotimes_{i=1}^{d} \gamma|_{\mathcal{X}_i}$ if and only if the cross cumulants vanish, that is $\kappa^{\mathbf{i}}(\gamma) = 0$ for all $\mathbf{i} \in \mathbb{N}_+^d$.*

Proof.
• Item 2: We want to show that the cross cumulants vanish if and only if $\gamma = \bigotimes_{i=1}^{d} \gamma|_{\mathcal{X}_i}$. By Lemma 2.C.1 it is enough to show that

$$\mathbb{E}_\gamma\Big[p\Big(\varphi_1(X_1), \ldots, \varphi_d(X_d)\Big)\Big] = \mathbb{E}_{\bigotimes_{i=1}^{d} \gamma|_{\mathcal{X}_i}}\Big[p\Big(\varphi_1(X_1), \ldots, \varphi_d(X_d)\Big)\Big]$$

for any monomial function p. Let us take linear functionals $F = (f_1, \ldots, f_d)$ and note that

$$\langle F^{\mathbf{i}}, \kappa^{\mathbf{i}}(\gamma) \rangle = \sum_{\pi \in P(d)} c_\pi \mathbb{E}_{\gamma_\pi^{\mathbf{i}}}\Big[f_1(\varphi_1(X_1)) \cdots f_d(\varphi_d(X_d))\Big]$$

which is the classical cumulant of the vector-valued random variable

$$\Big((f_1 \circ \varphi_1)(X_1), \ldots, (f_d \circ \varphi_d)(X_d)\Big),$$

where $(X_1, \ldots, X_d) \sim \gamma$. Hence by classical results [171], all cross cumulants of $\Big((f_1 \circ \varphi_1)(X_1), \ldots, (f_d \circ \varphi_d)(X_d)\Big)$ vanish if and only if the cross moments split, that is to say

$$\mathbb{E}_\gamma\Big[p\Big((f_1 \circ \varphi_1)(X_1), \ldots, (f_d \circ \varphi_d)(X_d)\Big)\Big] = \mathbb{E}_{\bigotimes_{i=1}^{d} \gamma|_{\mathcal{X}_i}}\Big[p\Big((f_1 \circ \varphi_1)(X_1), \ldots, (f_d \circ \varphi_d)(X_d)\Big)\Big]$$

for any monomial p on $\mathbb{R}^d$. Since $f_1, \ldots, f_d$ were arbitrary this holds for all monomials, which shows the assertion.

• Item 1: By assumption $\kappa^{\mathbf{i}}(\gamma) = \kappa^{\mathbf{i}}(\eta)$ for every $\mathbf{i} \in \mathbb{N}^d$; this implies that $\mathbb{E}_\gamma p(\varphi_1, \ldots, \varphi_d) = \mathbb{E}_\eta p(\varphi_1, \ldots, \varphi_d)$ for any polynomial p, so we can apply Lemma 2.C.1. □

Proposition 2.C.2 (Theorem 2.3.1 and Theorem 2.3.2)**.** *Let $\mathcal{X}_1, \ldots, \mathcal{X}_d$ be Polish spaces and $k_i : \mathcal{X}_i^2 \to \mathbb{R}$ be a collection of bounded, continuous, point-separating kernels. Let γ and η be be probability measures on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_d$. Then*

1. *$\gamma = \eta$ if and only if $\kappa_{k_1,\ldots,k_d}(\gamma) = \kappa_{k_1,\ldots,k_d}(\eta)$.*
2. *$\gamma = \bigotimes_{i=1}^d \gamma|_{\mathcal{X}_i}$ if and only if $\kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma) = 0$ for all $\mathbf{i} \in \mathbb{N}^d_+$.*

Proof. We reduce the proof to the checking of the conditions of Theorem 2.C.1. Let φ_i denote the canonical feature map of the kernel k_i, and let $\mathcal{B}_i := \mathcal{H}_{k_i}$ be the RKHS associated to k_i ($i \in \{1, \ldots, d\}$). For all $i \in \{1, \ldots, d\}$, φ_i is (i) bounded by the boundeness of k_i since $\|\varphi_i(x)\|^2_{\mathcal{H}_{k_i}} = k_i(x, x) \leq \sup_{x \in \mathcal{X}_i} |k_i(x, x)| < \infty$, (ii) continuous by the continuity of k_i [177, Lemma 4.29], (iii) injective by the point-separating property of k_i. The separability of $\mathcal{H}_{k_i}$ follows [177, Lemma 4.33] from the separability of $\mathcal{X}_i$ and the continuity of k_i ($i \in \{1, \ldots, d\}$). Note: Details on the expected kernel trick part of Theorem 2.3.1 and Theorem 2.3.2 are provided in Section 2.E. □

2.D Additional Experiments and Details

Here we give additional details on the experiments that were performed, and discuss some further experiments that did not fit into the main text.

Background on permutation testing. Permutation testing works by bootstrapping the distribution of a test statistic under the null hypothesis. This allows the user to estimate confidence intervals under the null, which is a powerful all-purpose way of doing so when analytic expressions are unavailable. As an example, assume we have two probability measures γ, η on $\mathcal{X}$ with i.i.d. samples $x_1, \ldots, x_N \sim \gamma, y_1, \ldots, y_N \sim \eta$. If the null hypothesis is that $\gamma = \eta$ then we may set

$$(z_1, \ldots, z_{2N}) := (x_1, \ldots, x_N, y_1, \ldots, y_N)$$

so that for any permutation σ on $2N$ elements, we get two different set of of i.i.d. samples from $\gamma = \eta$ by using the empirical measures

$$\tilde{\gamma}_\sigma := (z_{\sigma(1)}, \ldots, z_{\sigma(N)}), \quad \tilde{\eta}_\sigma := (z_{\sigma(N+1)}, \ldots, z_{\sigma(2N)})$$

and for any statistic $S : \mathcal{P}(\mathcal{X})^2 \to \mathbb{R}$, we may estimate $S(\gamma, \eta)$ under the null by sampling from $S(\tilde{\gamma}_\sigma, \tilde{\eta}_\sigma)$. If the null hypothesis were true, we might expect $S(\gamma, \eta)$ to lie in a region with high probability of the permutation estimator, and we can use this as a criteria for rejecting the null. Under fairly weak assumptions, this yields a test at the appropriate level [48].

Comparing a uniform and a mixture. Any uniform random variable over a symmetric interval will have 0 mean and skewness, so a symmetric mixture only needs to match the variance. If X is a 50/50 mixture of $U[a, b]$ and $U[-a, -b]$ then

$$\mathrm{Var}(X) = \frac{2}{3}\left(b^2 + ba + a^2\right)$$

so if Y is distributed according to $U[-c, c]$ then we only need to solve

$$b^2 + ba + a^2 = c^2$$

which is straightforward for a given a and c.

Computational complexity of estimators. The V-statistic for $d^{(2)}$ as written in Lemma 2.3.2 is bottlenecked by the matrix multiplications. We may note however that for two matrices $\mathbf{A}, \mathbf{B}$ it holds that

$$\mathrm{Tr}(\mathbf{A}^\top \mathbf{B}) = \langle \mathbf{A} \circ \mathbf{B} \rangle,$$

where $\langle \cdot \rangle$ denotes the sum over elements and $\circ$ denotes the Hadamard product. We also note that for for $\mathbf{H}_n = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ we have $\left(\mathbf{A}\mathbf{H}_n\right)_{i,j} = \frac{1}{n}\sum_{c=1}^n A_{i,c}$. Using both of these tricks we may compute both $d^{(2)}$ and CSIC without any matrix multiplications, which brings the computational complexity down to $O(N^2)$ for both. For a comparison of actual computation time, see Fig. 2.6 and Fig. 2.7, where the average computational times for out methods are compared to the KME and and HSIC for N between 50 and 2000.

Type I error on the Seoul Bicycle data. The results when comparing the winter data to itself is presented in Fig. 2.8. As we see the performance is similar for both estimators and lies between 5 and 10%.

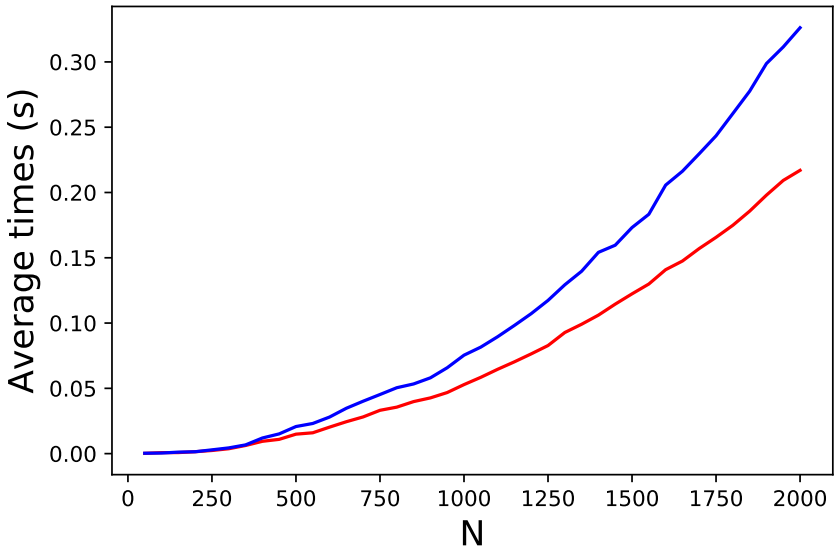


Figure 2.6: Average computational time in seconds for KME (red) and $d^{(2)}$ (blue) for sample size N between 50 and 2000.

Table 2.1: Comparison of classical and kernelized cumulants for independence testing with both variance and skewness.

N=20	Variance	Skewness
Classical	$19\% \pm 3.0\%$	$56\% \pm 3.5\%$
Rbf kernel	$39\% \pm 4.5\%$	$59\% \pm 3.0\%$

N=30	Variance	Skewness
Classical	$17\% \pm 0.5\%$	$68\% \pm 1.0\%$
Rbf kernel	$65\% \pm 3.5\%$	$79\% \pm 1.5\%$

Classical vs. kernelized cumulants. Using the same distributions as in the synthetic independence testing experiment, we now compare X with $Y_{0.5}^2$ to contrast independence testing with classical cumulants with their kernelized counterpart. The results are summarized in Table 2.1 where they are displayed as the median value $\pm$ half the difference between the 75th and 25th percentile. We consider every combination of classical vs. kernelized, variance vs. skewness, and two different sample sizes. One can observe that the classical variance based test performs poorly compared to a classical skewness test, the kernelized variance test is almost as powerful as the kernelized skewness test, and in all cases the kernelized tests deliver higher power.

2.E Kernel Trick Computations

Here we show how to arrive at the expressions used for the V-statistics used in the experiments.

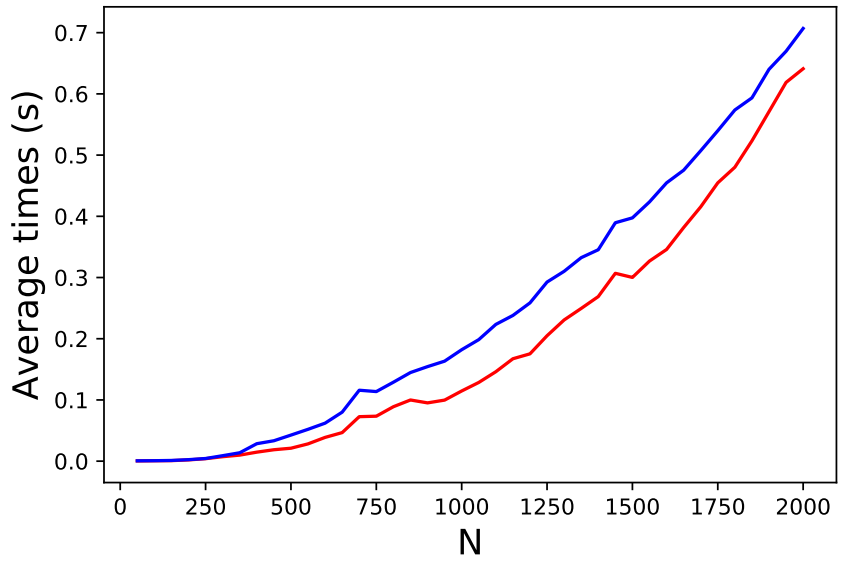


Figure 2.7: Average computational time in seconds for HSIC (red) and CSIC (blue) for sample size N between 50 and 2000.

Given a real analytic function $f(x,\dots,x_d)=\sum_{\mathbf{i}\in\mathbb{N}^d} f_{\mathbf{i}}x^{\mathbf{i}}$ in m variables with nonzero radius of convergence and Hilbert spaces $\mathcal{H}_1,\dots,\mathcal{H}_d$ we may (formally) extend f to a function

$$f_\otimes:\prod_{i=1}^{d}\mathcal{H}_i\to \mathrm{T},\quad f_\otimes(x_1,\dots,x_d)=\prod_{\mathbf{i}\in\mathbb{N}^d} f_{\mathbf{i}}x^{\otimes\mathbf{i}}.$$

Moreover, if the Hilbert spaces are RKHSs then we have the following result.

Lemma 2.E.1 (Nonlinear kernel trick)**.** *For any collection of RKHSs $\mathcal{H}_1,\dots,\mathcal{H}_d$ with feature maps $\varphi_i:\mathcal{X}_i\to\mathcal{H}_i$, assume that f and g are real analytic functions with radii of convergence $r(f)$ and $r(g)$ such that*

$$\max_{1\le i\le d}\sup_{x\in\mathcal{X}_i}|\varphi_i(x)|<\min(r(f),r(g)).$$

Then

$$\begin{aligned}&\langle f_\otimes\Big(\varphi_1(x_1),\dots,\varphi_d(x_d)\Big),g_\otimes\Big(\varphi_1(y_1),\dots,\varphi_d(y_d)\Big)\rangle_{\mathrm{T}}\\&=\sum_{\mathbf{i}\in\mathbb{N}^d} f_{\mathbf{i}}g_{\mathbf{i}}k_1(x_1,y_1)^{i_1}\dots k_d(x_d,y_d)^{i_d}.\end{aligned}$$

Proof. Since the image of the φ_is lie inside the radius of convergence of $f_\otimes$ and $g_\otimes$ the power series converge absolutely and we can write

$$\begin{aligned}&\langle f_\otimes\Big(\varphi^{\otimes\mathbf{i}}(x^{\mathbf{i}})\Big),g_\otimes\Big(\varphi^{\otimes\mathbf{i}}(y^{\mathbf{i}})\Big)\rangle_{\mathrm{T}}=\langle\sum_{\mathbf{i}\in\mathbb{N}^d} f_{\mathbf{i}}\varphi^{\otimes\mathbf{i}}(x^{\mathbf{i}}),\sum_{\mathbf{i}\in\mathbb{N}^d} g_{\mathbf{i}}\varphi^{\otimes\mathbf{i}}(y^{\mathbf{i}})\rangle_{\mathrm{T}}\\&=\sum_{\mathbf{i}\in\mathbb{N}^d} f_{\mathbf{i}}g_{\mathbf{i}}\langle\varphi^{\otimes\mathbf{i}}(x^{\mathbf{i}}),\varphi^{\otimes\mathbf{i}}(y^{\mathbf{i}})\rangle_{\mathcal{H}^{\otimes\mathbf{i}}}=\sum_{\mathbf{i}\in\mathbb{N}^d} f_{\mathbf{i}}g_{\mathbf{i}}k_1(x_1,y_1)^{i_1}\dots k_d(x_d,y_d)^{i_d},\end{aligned}$$

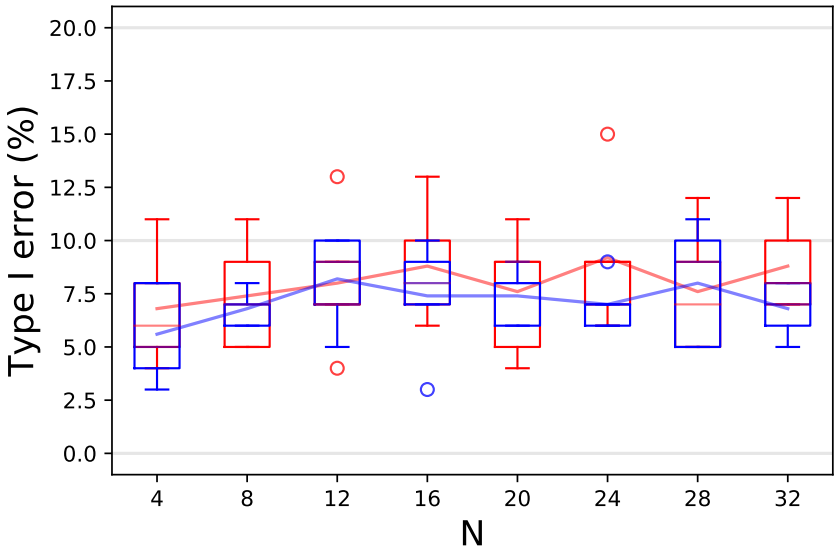


Figure 2.8: Type I errors using MMD (red) and $d^{(2)}$ (blue) on the Seoul bicycle data set.

where $\mathcal{H} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_d$. □

Using Lemma 2.E.1, we can choose kernels $k_i : \mathcal{X}_i^2 \to \mathbb{R}$ with associated RKHSs $\mathcal{H}_i$ and feature maps φ_i and some $\mathbf{i} \in \mathbb{N}^d$ with $\deg(\mathbf{i}) = m$. We make the observation that with $X = (X_1, \ldots, X_d) \sim \gamma$, $Y = (Y_1, \ldots, Y_d) \sim \eta$ and $k^{\otimes \mathbf{i}}$ and $\mathcal{H}^{\otimes \mathbf{i}}$ as in (2.4), one has

$$\begin{aligned}
&\langle \kappa^{\mathbf{i}}(\gamma), \kappa^{\mathbf{i}}(\eta) \rangle_{\mathcal{H}^{\otimes \mathbf{i}}} \\
&= \langle \sum_{\pi \in P(m)} c_\pi \mathbb{E}_{\gamma_\pi^{\mathbf{i}}} \varphi^{\otimes \mathbf{i}}(X^{\mathbf{i}}), \sum_{\tau \in P(m)} c_\tau \mathbb{E}_{\eta_\tau^{\mathbf{i}}} \varphi^{\otimes \mathbf{i}}(Y^{\mathbf{i}}) \rangle_{\mathcal{H}^{\otimes \mathbf{i}}} \\
&= \sum_{\pi, \tau \in P(m)} c_\pi c_\tau \langle \mathbb{E}_{\gamma_\pi^{\mathbf{i}}} \varphi^{\otimes \mathbf{i}}(X^{\mathbf{i}}), \mathbb{E}_{\eta_\tau^{\mathbf{i}}} \varphi^{\otimes \mathbf{i}}(Y^{\mathbf{i}}) \rangle_{\mathcal{H}^{\otimes \mathbf{i}}} \\
&= \sum_{\pi, \tau \in P(m)} c_\pi c_\tau \mathbb{E}_{\gamma_\pi^{\mathbf{i}} \otimes \eta_\tau^{\mathbf{i}}} \langle \varphi^{\otimes \mathbf{i}}(X^{\mathbf{i}}), \varphi^{\otimes \mathbf{i}}(Y^{\mathbf{i}}) \rangle_{\mathcal{H}^{\otimes \mathbf{i}}} \\
&= \sum_{\pi, \tau \in P(m)} c_\pi c_\tau \mathbb{E}_{\gamma_\pi^{\mathbf{i}} \otimes \eta_\tau^{\mathbf{i}}} k^{\otimes \mathbf{i}}((X_1, \ldots, X_m), (Y_1, \ldots, Y_m)),
\end{aligned}$$

Since

$$\begin{aligned}
\|\kappa^{\mathbf{i}}(\gamma)\|^2_{\mathcal{H}^{\otimes \mathbf{i}}} &= \langle \kappa^{\mathbf{i}}(\gamma), \kappa^{\mathbf{i}}(\gamma) \rangle_{\mathcal{H}^{\otimes \mathbf{i}}} \\
\|\kappa^{\mathbf{i}}(\gamma) - \kappa^{\mathbf{i}}(\eta)\|^2_{\mathcal{H}^{\otimes \mathbf{i}}} &= \langle \kappa^{\mathbf{i}}(\gamma), \kappa^{\mathbf{i}}(\gamma) \rangle_{\mathcal{H}^{\otimes \mathbf{i}}} + \langle \kappa^{\mathbf{i}}(\eta), \kappa^{\mathbf{i}}(\eta) \rangle_{\mathcal{H}^{\otimes \mathbf{i}}} - 2 \langle \kappa^{\mathbf{i}}(\gamma), \kappa^{\mathbf{i}}(\eta) \rangle_{\mathcal{H}^{\otimes \mathbf{i}}}
\end{aligned}$$

one gets the expected kernel trick statements of Theorem 2.3.1 and Theorem 2.3.2.

We are now interested in explicitly computing the expression $\|\kappa_{k,\ell}^{(1,2)}(\gamma)\|^2_{\mathcal{H}_k^{\otimes 1}\otimes\mathcal{H}_\ell^{\otimes 2}}$, $\|\kappa_k^{(2)}(\gamma)-\kappa_k^{(2)}(\eta)\|^2_{\mathcal{H}^{(1,1)}}$ and $\|\kappa_k^{(3)}(\gamma)-\kappa_k^{(3)}(\eta)\|^2_{\mathcal{H}_k^{\otimes 3}}$, and their corresponding V-statistics. Recall that for a (w.l.o.g.) symmetric, measurable function $h(z_1,\ldots,z_m)$, the V-statistic of h with N samples $Z_1,\ldots,Z_N$ is defined as

$$\mathrm{V}(h;Z_1,\ldots,Z_N) := N^{-m}\sum_{i_1,\ldots,i_m=1}^{N} h(Z_{i_1},\ldots,Z_{i_m}).$$

Under fairly general conditions, the V-statistic converges in distribution to $\mathbb{E}[h(Z_1,\ldots,Z_m)]$ and a well-developed theory describes this convergence [195, 168, 4].

Example 2.E.1 (Estimating $\|\kappa_k^{(2)}(\gamma)-\kappa_k^{(2)}(\eta)\|^2_{\mathcal{H}^{(1,1)}}$)**.** Let X,X',X'',X''' denote independent copies of γ and Y,Y',Y'',Y''' denote independent copies of η. The full expression for $\|\kappa_k^{(2)}(\gamma)-\kappa_k^{(2)}(\eta)\|^2_{\mathcal{H}^{(1,1)}}$ is

$$\begin{aligned}\|\kappa_k^{(2)}(\gamma)-\kappa_k^{(2)}(\eta)\|^2_{\mathcal{H}^{(1,1)}} =&\mathbb{E}k(X,X')k(X'',X''')+\mathbb{E}k(Y,Y')k(Y'',Y''')\\ &+\mathbb{E}k(X,X')^2+\mathbb{E}k(Y,Y')^2+2\mathbb{E}k(X,Y)k(X',Y)\\ &+2\mathbb{E}k(X,Y)k(X,Y')-2\mathbb{E}k(X,Y)k(X',Y')-2\mathbb{E}k(X,Y)^2\\ &-2\mathbb{E}k(X,X')k(X,X'')-2\mathbb{E}k(Y,Y')k(Y,Y'').\end{aligned}$$

Given samples $(x_i)_{i=1}^N$, $(y_i)_{i=1}^M$ from γ and η respectively the corresponding V statistic

is

$$
\begin{aligned}
&\frac{1}{N^4}\sum_{i,j,\kappa,l=1}^{N} k(x_i,x_j)k(x_\kappa,x_l) \\
+&\frac{1}{M^4}\sum_{i,j,\kappa,l=1}^{M} k(y_i,y_j)k(y_\kappa,y_l) \\
+&\frac{1}{N^2}\sum_{i,j=1}^{N} k(x_i,x_j)^2 \\
+&\frac{1}{M^2}\sum_{i,j=1}^{M} k(y_i,y_j)^2 \\
+&\frac{2}{N^2M}\sum_{i,\kappa=1}^{N}\sum_{j=1}^{M} k(x_i,y_j)k(x_\kappa,y_j) \\
+&\frac{2}{NM^2}\sum_{i=1}^{N}\sum_{j,\kappa=1}^{M} k(x_i,y_j)k(x_i,y_\kappa) \\
-&\frac{2}{N^2M^2}\sum_{i,l=1}^{N}\sum_{j,\kappa=1}^{M} k(x_i,y_j)k(x_\kappa,y_l) \\
-&\frac{2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} k(x_i,y_j)^2. \\
-&\frac{2}{N^3}\sum_{i,j,\kappa=1}^{N} k(x_i,x_j)k(x_i,x_\kappa) \\
-&\frac{2}{M^3}\sum_{i,j,\kappa=1}^{M} k(y_i,y_j)k(y_i,y_\kappa).
\end{aligned}
\tag{2.11}
$$

Let us define the Gram matrices $\mathbf{K}_x = [k(x_i,x_j)]_{i,j=1}^{N} \in \mathbb{R}^{N\times N}$, $\mathbf{K}_y = [k(y_i,y_j)]_{i,j=1}^{M} \in \mathbb{R}^{M\times M}$, $\mathbf{K}_{x,y} = [k(x_i,y_j)]_{i,j=1}^{N,M}$ and let $\mathbf{H}_N = \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top \in \mathbb{R}^{N\times N}$, $\mathbf{H}_M = \frac{1}{M}\mathbf{1}_M\mathbf{1}_M^\top \in \mathbb{R}^{M\times M}$ be the centering, then (2.11) can be rewritten as

$$
\begin{aligned}
&+\frac{1}{N^2}\mathrm{Tr}(\mathbf{H}_N\mathbf{K}_x\mathbf{H}_N\mathbf{K}_x) + \frac{1}{M^2}\mathrm{Tr}(\mathbf{H}_M\mathbf{K}_y\mathbf{H}_M\mathbf{K}_y) \\
&+\frac{1}{N^2}\mathrm{Tr}(\mathbf{K}_x^2) + \frac{1}{M^2}\mathrm{Tr}(\mathbf{K}_y^2) \\
&+\frac{2}{NM}\mathrm{Tr}(\mathbf{K}_{xy}\mathbf{H}_N\mathbf{K}_{xy}) + \frac{2}{NM}\mathrm{Tr}(\mathbf{K}_{xy}\mathbf{H}_M\mathbf{K}_{xy}^\top) \\
&-\frac{2}{NM}\mathrm{Tr}(\mathbf{H}_M\mathbf{K}_{xy}^\top\mathbf{H}_N\mathbf{K}_{xy}) - \frac{2}{NM}\mathrm{Tr}(\mathbf{K}_{xy}^2) \\
&-\frac{2}{N^2}\mathrm{Tr}(\mathbf{K}_x\mathbf{H}_N\mathbf{K}_x) - \frac{2}{M^2}\mathrm{Tr}(\mathbf{K}_y\mathbf{H}_M\mathbf{K}_y)
\end{aligned}
$$

which simplifies to

$$\frac{1}{N^2}\mathrm{Tr}\Big[(\mathbf{K}_x(\mathbf{I}-\mathbf{H}_N))^2\Big]+\frac{1}{M^2}\mathrm{Tr}\Big[(\mathbf{K}_y(\mathbf{I}-\mathbf{H}_M))^2\Big]$$
$$-\frac{2}{NM}\mathrm{Tr}\Big[\mathbf{K}_{xy}(\mathbf{I}-\mathbf{H}_M)\mathbf{K}_{xy}^\top(\mathbf{I}-\mathbf{H}_N)\Big].$$

This estimator can be computed in quadratic time.

Example 2.E.2 (Estimating $\|\kappa_{k,\ell}^{(1,2)}(\gamma)\|^2_{\mathcal{H}_k^{\otimes 1}\otimes\mathcal{H}_\ell^{\otimes 2}}$)**.** Let k denote the kernel on $\mathcal{X}_1$ and ℓ denote the kernel on $\mathcal{X}_2$. Let $(X,Y),(X',Y'),(X'',Y''),(X^{(3)},Y^{(3)}),(X^{(4)},Y^{(4)}),$ $(X^{(5)},Y^{(5)})$ denote independent copies of $\gamma\in\mathcal{P}(\mathcal{X}_1\times\mathcal{X}_2)$. The full expression for $\|\kappa_{k,\ell}^{(1,2)}(\gamma)\|^2_{\mathcal{H}_k^{\otimes 1}\otimes\mathcal{H}_\ell^{\otimes 2}}$ is

$$\begin{aligned}
&\mathbb{E}k(X,X')k(X,X')\ell(Y,Y')-4\mathbb{E}k(X,X')k(X,X'')\ell(Y,Y')\\
&-2\mathbb{E}k(X,X')k(X,X')\ell(Y,Y'')+4\mathbb{E}k(X,X')k(X,X'')\ell(Y,Y^{(3)})\\
&+2\mathbb{E}k(X,X')k(X'',X^{(3)})\ell(Y,Y')+2\mathbb{E}k(X,X')k(X'',X^{(3)})\ell(Y,Y^{(3)})\\
&+4\mathbb{E}k(X,X')k(X'',X')\ell(Y,Y^{(3)})+\mathbb{E}k(X,X')k(X,X')\ell(Y'',Y^{(3)})\\
&-8\mathbb{E}k(X,X')k(X'',X^{(3)})\ell(Y^{(4)},Y')-4\mathbb{E}k(X,X')k(X'',X')\ell(Y^{(4)},Y^{(3)})\\
&+4\mathbb{E}k(X,X')k(X'',X^{(3)})\ell(Y^{(4)},Y^{(5)}).
\end{aligned}$$

Given samples $(x_i, y_i)_{i=1}^N$ from γ the corresponding V-statistic for this expression is

$$\begin{aligned}
&\frac{1}{N^2}\sum_{i,j=1}^{N} k(x_i,x_j)k(x_i,x_j)\ell(y_i,y_j)\\
-&\frac{4}{N^3}\sum_{i,j,\kappa=1}^{N} k(x_i,x_j)k(x_i,x_\kappa)\ell(y_i,y_j)\\
-&\frac{2}{N^3}\sum_{i,j,\kappa=1}^{N} k(x_i,x_j)k(x_i,x_j)\ell(y_i,y_\kappa)\\
+&\frac{4}{N^4}\sum_{i,j,\kappa,l=1}^{N} k(x_i,x_j)k(x_i,x_\kappa)\ell(y_i,y_l)\\
+&\frac{2}{N^4}\sum_{i,j,\kappa,l=1}^{N} k(x_i,x_j)k(x_\kappa,x_l)\ell(y_i,y_j)\\
+&\frac{2}{N^4}\sum_{i,j,\kappa,l=1}^{N} k(x_i,x_j)k(x_\kappa,x_l)\ell(y_i,y_l)\\
+&\frac{4}{N^4}\sum_{i,j,\kappa,l=1}^{N} k(x_i,x_j)k(x_\kappa,x_j)\ell(y_i,y_l)\\
+&\frac{1}{N^4}\sum_{i,j,\kappa,l=1}^{N} k(x_i,x_j)k(x_i,x_j)\ell(y_\kappa,y_l)\\
-&\frac{8}{N^5}\sum_{i,j,\kappa,l,m=1}^{N} k(x_i,x_j)k(x_\kappa,x_l)\ell(y_m,y_j)\\
-&\frac{4}{N^5}\sum_{i,j,\kappa,l,m=1}^{N} k(x_i,x_j)k(x_\kappa,x_j)\ell(y_m,y_l)\\
+&\frac{4}{N^6}\sum_{i,j,\kappa,l,m,n=1}^{N} k(x_i,x_j)k(x_\kappa,x_l)\ell(y_m,y_n).
\end{aligned}$$

Using the shorthand notation $\mathbf{K} = \mathbf{K}_x, \mathbf{L} = \mathbf{L}_y$ and $\mathbf{H} = \mathbf{H}_N$ and denoting by $\circ$ the Hadamard product $[\mathbf{A} \circ \mathbf{B}]_{i,j} = A_{i,j}B_{i,j}$ and $\langle\cdot\rangle$ the sum over all elements of a matrix $\langle\mathbf{A}\rangle = \sum_{i,j=1}^N A_{i,j}$, the V-statistic above can be written in the simpler form

$$\begin{aligned}
&\frac{1}{N^2}\Bigg\langle \mathbf{K}\circ\mathbf{K}\circ\mathbf{L} - 4\mathbf{K}\circ\mathbf{KH}\circ\mathbf{L} - 2\mathbf{K}\circ\mathbf{K}\circ\mathbf{LH}\\
&+4\mathbf{KH}\circ\mathbf{K}\circ\mathbf{LH} + 2\mathbf{K}\circ\mathbf{L}\left\langle\frac{\mathbf{K}}{N^2}\right\rangle + 2\mathbf{KH}\circ\mathbf{HK}\circ\mathbf{L}\\
&+4\mathbf{K}\circ\mathbf{HK}\circ\mathbf{LH} + \mathbf{K}\circ\mathbf{K}\left\langle\frac{\mathbf{L}}{N^2}\right\rangle - 8\mathbf{K}\circ\mathbf{LH}\left\langle\frac{\mathbf{K}}{N^2}\right\rangle\\
&-4\mathbf{K}\circ\mathbf{HK}\left\langle\frac{\mathbf{L}}{N^2}\right\rangle + 4\left\langle\frac{\mathbf{K}}{N^2}\right\rangle^2\mathbf{L}\Bigg\rangle.
\end{aligned}$$

Again this estimator can be computed in quadratic time.

Example 2.E.3 (Estimating $\|\kappa_k^{(3)}(\gamma) - \kappa_k^{(3)}(\eta)\|^2_{\mathcal{H}_k^{\otimes 3}}$)**.** In order to estimate $d^{(3)}(\gamma, \eta)$ we note that one can write

$$\begin{aligned}\|\kappa_k^{(3)}(\gamma) - \kappa_k^{(3)}(\eta)\|^2_{\mathcal{H}_k^{\otimes 3}} &= \|\kappa_k^{(3)}(\gamma)\|^2_{\mathcal{H}_k^{\otimes 3}} + \|\kappa_k^{(3)}(\eta)\|^2_{\mathcal{H}_k^{\otimes 3}} \\ &\quad - 2\langle \kappa_k^{(3)}(\gamma), \kappa_k^{(3)}(\eta)\rangle_{\mathcal{H}_k^{\otimes 3}}.\end{aligned}$$

We can estimate the first two terms like in Example 2.E.2, and the third term can be expressed as

$$\begin{aligned}\langle \kappa_k^{(3)}(\gamma), \kappa_k^{(3)}(\eta)\rangle_{\mathcal{H}_k^{\otimes 3}} &= \mathbb{E}k(X,Y)^3 - 3\mathbb{E}k(X,Y)^2k(X,Y')^2 \\ &\quad -3\mathbb{E}k(X,Y)^2k(X',Y)^2 + 6\mathbb{E}k(X,Y)k(X,Y')k(X',Y) + 3\mathbb{E}k(X,Y)^2k(X',Y') \\ &\quad +2\mathbb{E}k(X,Y)k(X',Y)k(X'',Y) + 2\mathbb{E}k(X,Y)k(X,Y')k(X,Y'') \\ &\quad -6\mathbb{E}k(X,Y)k(X,Y')k(X',Y'') - 6\mathbb{E}k(X,Y)k(X',Y)k(X'',Y') \\ &\quad +4\mathbb{E}k(X,Y)k(X',Y')k(X'',Y'').\end{aligned}$$

For simplicity we will assume that we have an equal number of samples (N) from both measures $(x_i)_{i=1}^N \in \gamma$ and $(y_i)_{i=1}^N \in \eta$. The V-statistic for $\langle \kappa_k^{(3)}(\gamma), \kappa_k^{(3)}(\eta)\rangle_{\mathcal{H}_k^{\otimes 3}}$ can be expressed as

$$\begin{aligned}&\frac{1}{N^2}\sum_{i,j=1}^N k(x_i,y_j)^3 \\ -&\frac{3}{N^3}\sum_{i,j,\kappa=1}^N k(x_i,y_j)^2k(x_i,y_\kappa) \\ -&\frac{3}{N^3}\sum_{i,j,\kappa=1}^N k(x_i,y_j)^2k(x_\kappa,y_i) \\ +&\frac{6}{N^4}\sum_{i,j,\kappa,l=1}^N k(x_i,y_j)k(x_i,y_\kappa)k(x_l,y_j) \\ +&\frac{3}{N^4}\sum_{i,j,\kappa,l=1}^N k(x_i,y_j)^2k(x_\kappa,y_l) \\ +&\frac{2}{N^4}\sum_{i,j,\kappa,l=1}^N k(x_i,y_j)k(x_\kappa,y_j)k(x_l,y_j)\end{aligned}$$

$$
\begin{aligned}
&+\frac{2}{N^4}\sum_{i,j,\kappa,l=1}^{N} k(x_i,y_j)k(x_i,y_\kappa)k(x_i,y_l)\\
&-\frac{6}{N^5}\sum_{i,j,\kappa,l,m=1}^{N} k(x_i,y_j)k(x_i,y_\kappa)k(x_l,y_m)\\
&-\frac{6}{N^5}\sum_{i,j,\kappa,l,m=1}^{N} k(x_i,y_j)k(x_\kappa,y_j)k(x_l,y_m)\\
&+\frac{4}{N^6}\sum_{i,j,\kappa,l,m,n=1}^{N} k(x_i,y_j)k(x_\kappa,y_l)k(x_m,y_n).
\end{aligned}
$$

Using the notation $\mathbf{K}_{xy} = [k(x_i, y_j)]_{i,j=1}^N$, this estimator simplifies to

$$
\begin{aligned}
&\frac{1}{N^2}\Big\langle \mathbf{K}_{xy}\circ\mathbf{K}_{xy}\circ\mathbf{K}_{xy} - 3\mathbf{K}_{xy}\circ\mathbf{K}_{xy}\circ\mathbf{H}\mathbf{K}_{xy}\\
&-3\mathbf{K}_{xy}\circ\mathbf{K}_{xy}\circ\mathbf{K}_{xy}\mathbf{H} + 6\mathbf{K}_{xy}\circ\mathbf{K}_{xy}\mathbf{H}\circ\mathbf{H}\mathbf{K}_{xy}\\
&+3\mathbf{K}_{xy}\circ\mathbf{K}_{xy}\left\langle\frac{\mathbf{K}_{xy}}{N^2}\right\rangle + 2\mathbf{K}_{xy}\circ\mathbf{H}\mathbf{K}_{xy}\circ\mathbf{H}\mathbf{K}_{xy}\\
&+2\mathbf{K}_{xy}\circ\mathbf{K}_{xy}\mathbf{H}\circ\mathbf{K}_{xy}\mathbf{H} - 6\mathbf{K}_{xy}\circ\mathbf{K}_{xy}\mathbf{H}\left\langle\frac{\mathbf{K}_{xy}}{N^2}\right\rangle\\
&-6\mathbf{K}_{xy}\circ\mathbf{H}\mathbf{K}_{xy}\left\langle\frac{\mathbf{K}_{xy}}{N^2}\right\rangle + 4\left\langle\frac{\mathbf{K}}{N^2}\right\rangle^2\mathbf{K}_{xy}\Big\rangle.
\end{aligned}
$$

We mention also that the first two terms $\|\kappa_k^{(3)}(\gamma)\|^2_{\mathcal{H}_k^{\otimes 3}}, \|\kappa_k^{(3)}(\eta)\|^2_{\mathcal{H}_k^{\otimes 3}}$ can be computed a little more simply than in Example 2.E.2 since the expressions have more symmetry, using the notation $\mathbf{K}_x = [k(x_i, x_j)]_{i,j=1}^N$ we can write down the V-statistic for $\|\kappa_k^{(3)}(\gamma)\|^2_{\mathcal{H}_k^{\otimes 3}}$ as

$$
\begin{aligned}
&\frac{1}{N^2}\Big\langle \mathbf{K}_x\circ\mathbf{K}_x\circ\mathbf{K}_x - 6\mathbf{K}_x\circ\mathbf{K}_x\mathbf{H}\circ\mathbf{K}_x\\
&+4\mathbf{K}_x\mathbf{H}\circ\mathbf{K}_x\circ\mathbf{K}_x\mathbf{H} + 3\mathbf{K}_x\circ\mathbf{K}_x\left\langle\frac{\mathbf{K}_x}{N^2}\right\rangle\\
&+6\mathbf{K}_x\mathbf{H}\circ\mathbf{H}\mathbf{K}_x\circ\mathbf{K}_x - 12\mathbf{K}_x\circ\mathbf{H}\mathbf{K}_x\left\langle\frac{\mathbf{K}_x}{N^2}\right\rangle\\
&+4\left\langle\frac{\mathbf{K}_x}{N^2}\right\rangle^2\mathbf{K}_x\Big\rangle
\end{aligned}
$$

with a similar expression for $\|\kappa_k^{(3)}(\eta)\|^2_{\mathcal{H}_k^{\otimes 3}}$. The estimator can be computed in quadratic time.

Chapter 3

Seq2Tens: An Efficient Representation of Sequences by Low-Rank Tensor Projections

This chapter is based on [189] which was written with Csaba Toth and Harald Oberhauser. The main text was written by Csaba, but it is provided here for context. I wrote Appendix 3.A, 3.B, and 3.C. The other appendices from the full paper were left out.

Sequential data such as time series, video, or text can be challenging to analyse as the ordered structure gives rise to complex dependencies. At the heart of this is non-commutativity, in the sense that reordering the elements of a sequence can completely change its meaning. We use a classical mathematical object – the free algebra – to capture this non-commutativity. To address the innate computational complexity of this algebra, we use compositions of low-rank tensor projections. This yields modular and scalable building blocks that give state-of-the-art performance on standard benchmarks such as multivariate time series classification, mortality prediction and generative models for video. Code and benchmarks are publicly available at https://github.com/tgcsaba/seq2tens.

3.1 Introduction

A central task of learning is to find representations of the underlying data that efficiently and faithfully capture their structure. In the case of sequential data, one data point consists of a sequence of objects. This is a rich and non-homogeneous class of data and includes classical uni- or multi-variate time series (sequences of scalars or vectors), video (sequences of images), and text (sequences of letters). Particular challenges of sequential data are that each sequence entry can itself be a highly structured object and that data sets typically include sequences of different length which makes naive vectorization troublesome.

Contribution. Our main result is a generic method that takes a *static feature map* for a class of objects (e.g. a feature map for vectors, images, or letters) as input and turns this into a feature map for sequences of arbitrary length of such objects (e.g. a feature map for time series, video, or text). We call this feature map for sequences Seq2Tens for reasons that will become clear; among its attractive properties are that it

(i) provides a structured, parsimonious description of sequences; generalizing classical methods for strings,

(ii) comes with theoretical guarantees such as universality,

(iii) can be turned into modular and flexible neural network (NN) layers for sequence data.

The key ingredient to our approach is to embed the feature space of the static feature map into a larger linear space that forms an algebra (a vector space equipped with a multiplication). The product in this algebra is then used to "stitch together" the static features of the individual sequence entries in a structured way. The construction that allows to do all this is classical in mathematics, and known as the *free algebra* (over the static feature space).

Outline. Section 3.2 formalizes the main ideas of Seq2Tens and introduces the free algebra $\mathrm{T}(V)$ over a space V as well as the associated product, the so-called *convolution tensor product*. Section 3.3 shows how low rank (LR) constructions combined with sequence-to-sequence transforms allows one to efficiently use this rich algebraic structure. Section 3.4 applies the results of Sections 3.2 and 3.3 to build

modular and scalable NN layers for sequential data. Section 3.5 demonstrates the flexibility and modularity of this approach on both discriminative and generative benchmarks. Section 3.6 makes connections with previous work and summarizes this article. In the appendices we provide mathematical background, extensions, and detailed proofs for our theoretical results.

3.2 Capturing order by non-commutative multiplication

We denote the set of sequences of elements in a set $\mathcal{X}$ by

$$\operatorname{Seq} \mathcal{X} = \{\mathbf{x} = (\mathbf{x}_i)_{i=1,\dots,L} : \mathbf{x}_i \in \mathcal{X},\, L \geq 1\}$$

where $L \geq 1$ is some arbitrary length. Even if $\mathcal{X}$ itself is a linear space, e.g. $\mathcal{X} = \mathbb{R}$, $\operatorname{Seq} \mathcal{X}$ is never a linear space since there is no natural addition of two sequences of different length.

Seq2Tens in a nutshell. Given any vector space V we may construct the so-called free algebra $\mathrm{T}(V)$ over V. We describe the space $\mathrm{T}(V)$ in detail below, but as for now the only thing that is important is that $\mathrm{T}(V)$ is also a vector space that includes V, and that it carries a non-commutative product, which is, in a precise sense, "the most general product" on V.

The main idea of Seq2Tens is that any "static feature map" for elements in $\mathcal{X}$

$$\phi : \mathcal{X} \to V$$

can be used to construct a new feature map $\Phi : \operatorname{Seq} \mathcal{X} \to \mathrm{T}(V)$ for sequences in $\mathcal{X}$ by using the algebraic structure of $\mathrm{T}(V)$: the non-commutative product on $\mathrm{T}(V)$ makes it possible to "stitch together" the individual features $\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_L) \in V \subset \mathrm{T}(V)$ of the sequence $\mathbf{x}$ in the larger space $\mathrm{T}(V)$ by multiplication in $\mathrm{T}(V)$. With this we may define the feature map $\Phi(\mathbf{x})$ for a sequences $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_L) \in \operatorname{Seq} \mathcal{X}$ as follows

(i) lift the map $\phi : \mathcal{X} \to V$ to a map $\varphi : \mathcal{X} \to \mathrm{T}(V)$,

(ii) map $\operatorname{Seq} \mathcal{X} \to \operatorname{Seq} \mathrm{T}(V)$ by $(\mathbf{x}_1, \dots, \mathbf{x}_L) \mapsto (\varphi(\mathbf{x}_1), \dots, \varphi(\mathbf{x}_L))$,

(iii) map $\operatorname{Seq} \mathrm{T}(V) \to \mathrm{T}(V)$ by multiplication $(\varphi(\mathbf{x}_1), \dots, \varphi(\mathbf{x}_L)) \mapsto \varphi(\mathbf{x}_1) \cdots \varphi(\mathbf{x}_L)$.

In a more concise form, we define Φ as

$$\Phi : \operatorname{Seq} \mathcal{X} \to \mathrm{T}(V), \quad \Phi(\mathbf{x}) = \prod_{i=1}^{L} \varphi(\mathbf{x}_i) \tag{3.1}$$

where $\prod$ denotes multiplication in $\mathrm{T}(V)$. We refer to the resulting map Φ as the Seq2Tens map, which stands short for ***Seq****uences-****2****-****Tens****ors*. Why is this construction a good idea? First note, that step (i) is always possible since $V \subset \mathrm{T}(V)$ and we discuss the simplest such lift before Theorem 3.2.1 as well as other choices in Appendix 3.B. Further, if ϕ, respectively φ, provides a faithful representation of objects in $\mathcal{X}$, then there is no loss of information in step (ii). Finally, since step (iii) uses "the most general product" to multiply $\varphi(\mathbf{x}_1) \cdots \varphi(\mathbf{x}_L)$ one expects that $\Phi(\mathbf{x}) \in \mathrm{T}(V)$ faithfully represents the sequence $\mathbf{x}$ as an element of $\mathrm{T}(V)$. Indeed in Theorem 3.2.1 below we show an even stronger statement, namely that if the static feature map $\phi : \mathcal{X} \to V$ contains enough non-linearities so that non-linear functions from $\mathcal{X}$ to $\mathbb{R}$ can be approximated as *linear functions* of the static feature map ϕ, then the above construction extends this property to functions of sequences. Put differently, *if ϕ is a universal feature map for $\mathcal{X}$, then Φ is a universal feature map for* $\mathrm{Seq}\,\mathcal{X}$; that is, any non-linear function $f(\mathbf{x})$ of a sequence $\mathbf{x}$ can be approximated as a linear functional of $\Phi(\mathbf{x})$, $f(\mathbf{x}) \approx \langle \ell, \Phi(\mathbf{x}) \rangle$. We also emphasize that the domain of Φ is the space $\mathrm{Seq}\,\mathcal{X}$ of sequences of *arbitrary* (finite) length. The remainder of this Section gives more details about steps (i),(ii),(iii) for the construction of Φ.

The free algebra $\mathrm{T}(V)$ over a vector space V. Let V be a vector space. We denote by $\mathrm{T}(V)$ the set of sequences of tensors indexed by their degree m,

$$\mathrm{T}(V) := \{\mathbf{t} = (\mathbf{t}_m)_{m \geq 0} \,|\, \mathbf{t}_m \in V^{\otimes m}\}$$

where by convention $V^{\otimes 0} = \mathbb{R}$. For example, if $V = \mathbb{R}^d$ and $\mathbf{t} = (\mathbf{t}_m)_{m \geq 0}$ is some element of $\mathrm{T}(\mathbb{R}^d)$, then its degree $m = 1$ component is a d-dimensional vector $\mathbf{t}_1$, its degree $m = 2$ component is a $d \times d$ matrix $\mathbf{t}_2$, and its degree $m = 3$ component is a degree 3 tensor $\mathbf{t}_3$. By defining addition and scalar multiplication as

$$\mathbf{s} + \mathbf{t} := (\mathbf{s}_m + \mathbf{t}_m)_{m \geq 0}, \quad c \cdot \mathbf{t} = (c\mathbf{t}_m)_{m \geq 0}$$

the set $\mathrm{T}(V)$ becomes a linear space. By identifying $v \in V$ as the element $(0, v, 0, \ldots, 0) \in \mathrm{T}(V)$ we see that V is a linear subspace of $\mathrm{T}(V)$. Moreover, while V is only a linear space, $\mathrm{T}(V)$ carries a product that turns $\mathrm{T}(V)$ into an algebra.

This product is the so-called *tensor convolution product*, and is defined for $\mathbf{s}, \mathbf{t} \in \mathrm{T}(V)$ as

$$\mathbf{s} \cdot \mathbf{t} := \Big(\sum_{i=0}^{m} \mathbf{s}_i \otimes \mathbf{t}_{m-i} \Big)_{m \geq 0} = \Big(1, \mathbf{s}_1 + \mathbf{t}_1, \mathbf{s}_2 + \mathbf{s}_1 \otimes \mathbf{t}_1 + \mathbf{t}_2, \ldots \Big) \in \mathrm{T}(V) \quad (3.2)$$

where $\otimes$ denotes the usual outer tensor product; e.g. for vectors $u = (u_i), v = (v_i) \in \mathbb{R}^d$ the outer tensor product $u \otimes v$ is the $d \times d$ matrix $(u_i v_j)_{i,j=1,\dots,d}$. We emphasize that like the outer tensor product $\otimes$, the tensor convolution product $\cdot$ is non-commutative, i.e. $\mathbf{s} \cdot \mathbf{t} \neq \mathbf{t} \cdot \mathbf{s}$. In a mathematically precise sense, $\mathrm{T}(V)$ *is the most general algebra that contains* V*; it is a "free construction".* Since $\mathrm{T}(V)$ is realized as series of tensors of increasing degree, the *free algebra* $\mathrm{T}(V)$ is also known as the *tensor algebra* in the literature. Appendix 3.A contains background on tensors and further examples.

Lifting static feature maps. Step (i) in the construction of Φ requires turning a given feature map $\phi : \mathcal{X} \to V$ into a map $\varphi : \mathcal{X} \to \mathrm{T}(V)$. Throughout the rest of this article we use the lift

$$\varphi(\mathbf{x}) = (1, \phi(\mathbf{x}), 0, 0 \dots) \in \mathrm{T}(V). \tag{3.3}$$

We discuss other choices in Appendix 3.B, but attractive properties of the lift 3.3 are that

(a) the evaluation of Φ against low rank tensors becomes a simple recursive formula (Proposition 3.3.1,

(b) it is a generalization of sequence sub-pattern matching as used in string kernels (Appendix 3.B.3,

(c) despite its simplicity it performs exceedingly well in practice (Section 3.4).

Extending to sequences of arbitrary length. Steps (i) and (ii) in the construction specify how the map $\Phi : \mathcal{X} \to \mathrm{T}(V)$ behaves on sequences of length-1, that is, single observations. Step (iii) amounts to the requirement that for any two sequences $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_K), \mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_L) \in \operatorname{Seq} V$, their concatenation defined as $\mathbf{z} = (\mathbf{x}_1, \dots, \mathbf{x}_K, \mathbf{y}_1, \dots, \mathbf{y}_L) \in \operatorname{Seq} V$ can be understood in the feature space as (non-commutative) multiplication of their corresponding features

$$\Phi(\mathbf{z}) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}).$$

In other words, we inductively extend the lift φ to sequences of arbitrary length by starting from sequences consisting of a single observation, which is given in (3.1). Repeatedly applying the definition of the tensor convolution product in (3.2) leads to the following explicit formula

$$\Phi_m(\mathbf{x}) = \sum_{1 \le i_1 < \cdots < i_m \le L} \mathbf{x}_{i_1} \otimes \cdots \otimes \mathbf{x}_{i_m} \in V^{\otimes m}, \;\; \Phi(\mathbf{x}) = (\Phi_m(\mathbf{x}))_{m \ge 0}, \qquad (3.4)$$

where $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_L) \in \operatorname{Seq} V$ and the summation is over non-contiguous subsequences of $\mathbf{x}$.

Some intuition: generalized pattern matching. Our derivation of the feature map $\Phi(\mathbf{x}) = (1, \Phi_1(\mathbf{x}), \Phi_2(\mathbf{x}), \ldots) \in \mathrm{T}(V)$ was guided by general algebraic principles, but (3.4) provides an intuitive interpretation. It shows that for each $m \ge 1$, the entry $\Phi_m(\mathbf{x}) \in V^{\otimes m}$ constructs a summary of a long sequence $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_L) \in \operatorname{Seq} V$ based on subsequences $(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_m})$ of $\mathbf{x}$ of length-m. It does this by taking the usual outer tensor product $\mathbf{x}_{i_1} \otimes \cdots \otimes \mathbf{x}_{i_m} \in V^{\otimes m}$ and summing over all possible subsequences. This is completely analogous to how string kernels provide a structured description of text by looking at non-contiguous substrings of length-m (indeed, Appendix 3.B.3 makes this rigorous). However, the main difference is that the above construction works for arbitrary sequences and not just sequences of discrete letters. Readers with less mathematical background might simply take this as motivation and regard (3.4) as definition. However, the algebraic background allows to prove that Φ is universal, see Theorem 3.2.1 below.

Universality. A function $\phi : \mathcal{X} \to V$ is said to be *universal for* $\mathcal{X}$ if all continuous functions on $\mathcal{X}$ can be approximated as linear functions on the image of ϕ. One of the most powerful features of neural nets is their universality [103]. A very attractive property of Φ is that it preserves universality: if $\phi : \mathcal{X} \to V$ is universal for $\mathcal{X}$, then $\Phi : \operatorname{Seq} X \to \mathrm{T}(V)$ is universal for $\operatorname{Seq} \mathcal{X}$. To make this precise, note that $V^{\otimes m}$ is a linear space and therefore any $\ell = (\ell_0, \ell_1, \ldots, \ell_M, 0, 0, \ldots) \in \mathrm{T}(V)$ consisting of M tensors $\ell_m \in V^{\otimes m}$, yields a linear functional on $\mathrm{T}(V)$; e.g. if $V = \mathbb{R}^d$ and we identify ℓ_m in coordinates as $\ell_m = (\ell_m^{i_1,\ldots,i_m})_{i_1,\ldots,i_m \in \{1,\ldots,d\}}$ then

$$\langle \ell, \mathbf{t} \rangle := \sum_{m=0}^{M} \langle \ell_m, \mathbf{t}_m \rangle = \sum_{m=0}^{M} \sum_{i_1,\ldots,i_m \in \{1,\ldots,d\}} \ell_m^{i_1,\ldots,i_m} \mathbf{t}_m^{i_1,\ldots,i_m}.$$

Thus linear functionals of the feature map Φ, are real-valued functions of sequences. Theorem 3.2.1 below shows that any continuous function $f : \operatorname{Seq} \mathcal{X} \to \mathbb{R}$ can by arbitrary well approximated by a $\ell \in \mathrm{T}(V)$, $f(\mathbf{x}) \approx \langle \ell, \Phi(\mathbf{x}) \rangle$.

Theorem 3.2.1. *Let $\phi : \mathcal{X} \to V$ be a universal map with a lift that satisfies some mild constraints, then the following map is universal:*

$$\Phi : \mathrm{Seq}(\mathcal{X}) \to \mathrm{T}(V), \quad \mathbf{x} \mapsto \Phi(\mathbf{x}).$$

A detailed proof and the precise statement of Theorem 3.2.1 is given in Appendix 3.B.

3.3 Approximation by low-rank linear functionals

The combinatorial explosion of tensor coordinates and what to do about it. The universality of Φ suggests the following approach to represent a function $f : \mathrm{Seq}\,\mathcal{X} \to \mathbb{R}$ of sequences: First compute $\Phi(\mathbf{x})$ and then optimize over ℓ (and possibly also the hyperparameters of ϕ) such that $f(\mathbf{x}) \approx \langle \ell, \Phi(\mathbf{x}) \rangle = \sum_{m=0}^{M} \langle \ell_m, \Phi_m(\mathbf{x}) \rangle$. Unfortunately, tensors suffer from a combinatorial explosion in complexity in the sense that even just storing $\Phi_m(\mathbf{x}) \in V^{\otimes m} \subset \mathrm{T}(V)$ requires $O(\dim(V)^m)$ real numbers. Below we resolve this computational bottleneck as follows: in Proposition 3.3.1 we show that for a special class of low-rank elements $\ell \in \mathrm{T}(V)$, the functional $\mathbf{x} \mapsto \langle \ell, \Phi(\mathbf{x}) \rangle$ can be efficiently computed in both time and memory. This is somewhat analogous to a kernel trick since it shows that $\langle \ell, \Phi(\mathbf{x}) \rangle$ can be cheaply computed without explicitly computing the feature map $\Phi(\mathbf{x})$. However, Theorem 3.2.1 guarantees universality under no restriction on ℓ, thus restriction to rank-1 functionals limits the class of functions $f(\mathbf{x})$ that can be approximated. Nevertheless, by iterating these "low-rank functional" constructions in the form of sequence-to-sequence transformations this can be ameliorated. We give the details below but to gain intuition, we invite the reader to think of this iteration analogous to stacking layers in a neural network: each layer is a relatively simple non-linearity (e.g. a sigmoid composed with an affine function) but by composing such layers, complicated functions can be efficiently approximated.

Rank-1 functionals are computationally cheap. Degree $m = 2$ tensors are matrices and low-rank (LR) approximations of matrices are widely used in practice [194] to address the quadratic complexity. The definition below generalizes the rank of matrices (tensors of degree $m = 2$) to tensors of any degree m.

Definition 12. The *rank* (also called *CP rank* [36]) of a degree-m tensor $\mathbf{t}_m \in V^{\otimes m}$ is the smallest number $r \geq 0$ such that one may write

$$\mathbf{t}_m = \sum_{i=0}^{r} \mathbf{v}_i^1 \otimes \cdots \otimes \mathbf{v}_i^m, \quad \mathbf{v}_i^1, \ldots, \mathbf{v}_i^m \in V.$$

We say that $\mathbf{t} = (\mathbf{t}_m)_{m\geq 0} \in \mathrm{T}(V)$ has rank-1 (and degree-M) if each $\mathbf{t}_m \in V^{\otimes m}$ is a rank-1 tensor and $\mathbf{t}_i = 0$ for $i > M$.

Remark 3.3.1. For $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_L) \in \operatorname{Seq} V$, the rank $r_m \in \mathbb{N}$ of $\Phi_m(\mathbf{x})$ satisfies $r_m \leq \binom{L}{m}$, while the rank and degree $r, d \in \mathbb{N}$ of $\Phi(\mathbf{x})$ satisfy $r \leq \binom{L}{K}$ for $K = \left\lfloor \frac{L}{2} \right\rfloor$ and $d \leq L$.

A direct calculation shows that if ℓ is of rank-1, then $\langle \ell, \Phi(\mathbf{x}) \rangle$ can be computed very efficiently by inner product evaluations in V.

Proposition 3.3.1. *Let $\ell = (\ell_m)_{m\geq 0} \in \mathrm{T}(V)$ be of rank-1 and degree-M. If ϕ is lifted to φ as in* (3.3)*, then*

$$\langle \ell, \Phi(\mathbf{x}) \rangle = \sum_{m=0}^{M} \sum_{1 \leq i_1 < \cdots < i_m \leq L} \prod_{k=1}^{m} \langle \mathbf{v}_k^m, \phi(\mathbf{x}_{i_k}) \rangle \tag{3.5}$$

where $\ell_m = \mathbf{v}_1^m \otimes \cdots \otimes \mathbf{v}_m^m \in V^{\otimes m}$, $\mathbf{v}_i^m \in V$ and $m = 0, \ldots, M$.

Note that the inner sum is taken over all non-contiguous subsequences of $\mathbf{x}$ of length-m, analogously to m-mers of strings and we make this connection precise in Appendix 3.B.3; the proof of Proposition 3.3.1 is given in Appendix 3.B.1. While (3.5) looks expensive, by casting it into a recursive formulation over time, it can be computed in $O(M^2 \cdot L \cdot d)$ time and $O(M^2 \cdot (L + c))$ memory, where d is the inner product evaluation time on V, while c is the memory footprint of a $v \in V$. This can further be reduced to $O(M \cdot L \cdot d)$ time and $O(M \cdot (L + c))$ memory by an efficient parametrization of the rank-1 element $\ell \in \mathrm{T}(V)$. We give further details in [189, Appendices D.2, D.3, D.4].

Low-rank Seq2Tens maps. The composition of a linear map $\mathcal{L} : \mathrm{T}(V) \to \mathbb{R}^N$ with Φ can be computed cheaply in parallel using (3.5) when $\mathcal{L}$ is specified through a collection of $N \in \mathbb{N}$ rank-1 elements $\ell^1, \ldots, \ell^N \in \mathrm{T}(V)$ such that

$$\tilde{\Phi}_{\tilde{\theta}}(\mathbf{x}_1, \ldots, \mathbf{x}_L) := \mathcal{L} \circ \Phi(\mathbf{x}_1, \ldots, \mathbf{x}_L) = (\langle \ell^j, \Phi(\mathbf{x}_1, \ldots, \mathbf{x}_L) \rangle)_{j=1}^N \in \mathbb{R}^N.$$

We call the resulting map $\tilde{\Phi}_{\tilde{\theta}} : \operatorname{Seq} \mathcal{X} \to \mathbb{R}^N$ a ***L**ow-rank **S**eq**2**-**T**ens* map of width-N and order-M, where $M \in \mathbb{N}$ is the maximal degree of $\ell^1, \ldots, \ell^N$ such that $\ell_i^j = 0$ for $i > M$. The LS2T map is parametrized by

(1) the component vectors $\mathbf{v}_{j,m}^k \in V$ of the rank-1 elements $\ell_m^j = \mathbf{v}_{j,m}^1 \otimes \cdots \otimes \mathbf{v}_{j,m}^m$,

(2) by any parameters θ that the static feature map $\phi_\theta : \mathcal{X} \to V$ may depend on. We jointly denote these parameters by $\tilde{\theta} = (\theta, \ell^1, \ldots, \ell^N)$

. In addition, by the subsequent composition of $\tilde{\Phi}_{\tilde{\theta}}$ with a linear functional $\mathbb{R}^N \to \mathbb{R}$, we get the following function subspace as hypothesis class for the LS2T

$$\tilde{\mathcal{H}} = \Big\{ \langle \sum_{j=1}^{N} \alpha_j \ell^j, \Phi(\mathbf{x}_1, \dots, \mathbf{x}_L) \rangle \,|\, \alpha_j \in \mathbb{R} \Big\} \subsetneq \mathcal{H} = \Big\{ \langle \ell, \Phi(\mathbf{x}_1, \dots, \mathbf{x}_L) \,|\, \ell \in \mathrm{T}(V) \Big\}$$

Hence, we acquire an intuitive explanation of the (hyper)parameters: the width of the LS2T, $N \in \mathbb{N}$ specifies the maximal rank of the low-rank linear functionals of Φ that the LS2T can represent, while the span of the rank-1 elements, $\mathrm{span}(\ell^1, \dots, \ell^N)$ determine an N-dimensional subspace of the dual space of $\mathrm{T}(V)$ consisting of at most rank-N functionals.

Recall now that without rank restrictions on the linear functionals of Seq2Tens features, Theorem 3.2.1 would guarantee that any real-valued function $f : \mathrm{Seq}\,\mathcal{X} \to \mathbb{R}$ could be approximated by $f(\mathbf{x}) \approx \langle \ell, \Phi(\mathbf{x}_1, \dots, \mathbf{x}_L) \rangle$. As pointed out before, the restriction of the hypothesis class to low-rank linear functionals of $\Phi(\mathbf{x}_1, \dots, \mathbf{x}_L)$ would limit the class of functions of sequences that can be approximated. To ameliorate this, we use LS2T transforms in a sequence-to-sequence fashion that allows us to stack such low-rank functionals, significantly recovering expressiveness.

Sequence-to-sequence transforms. We can use LS2T to build sequence-to-sequence transformations in the following way: fix the static map $\phi_\theta : \mathcal{X} \to V$ parametrized by θ and rank-1 elements such that $\tilde{\theta} = (\theta, \ell^1, \dots, \ell^N)$ and apply the resulting LS2T map $\tilde{\Phi}_{\tilde{\theta}}$ over expanding windows of $\mathbf{x}$:

$$\mathrm{Seq}\,\mathcal{X} \to \mathrm{Seq}\,\mathbb{R}^N, \quad \mathbf{x} \mapsto \Big(\tilde{\Phi}_{\tilde{\theta}}(\mathbf{x}_1), \tilde{\Phi}_{\tilde{\theta}}(\mathbf{x}_1, \mathbf{x}_2), \dots, \tilde{\Phi}_{\tilde{\theta}}(\mathbf{x}_1, \dots, \mathbf{x}_L) \Big). \tag{3.6}$$

Note that the cost of computing the expanding window sequence-to-sequence transform in (3.6) is no more expensive than computing $\tilde{\Phi}_{\tilde{\theta}}(\mathbf{x}_1, \dots, \mathbf{x}_L)$ itself due to the recursive nature of our algorithms, for further details see [189, Appendix D.2, D.3, D.4].

Deep sequence-to-sequence transforms. Inspired by the empirical successes of deep RNNs [89, 88, 182], we iterate the transformation 3.6 D-times:

$$\mathrm{Seq}\,\mathcal{X} \to \mathrm{Seq}\,\mathbb{R}^{N_1} \to \mathrm{Seq}\,\mathbb{R}^{N_2} \to \cdots \to \mathrm{Seq}\,\mathbb{R}^{N_D}.$$

Each of these mappings $\mathrm{Seq}\,\mathbb{R}^{N_i} \to \mathrm{Seq}\,\mathbb{R}^{N_{i+1}}$ is parametrized by the parameters $\tilde{\theta}_i$ of a static feature map ϕ_{θ_i} and a linear map $\mathcal{L}_i$ specified by N_i rank-1 elements of $\mathrm{T}(V)$; these parameters are collectively denoted by $\tilde{\theta}_i = (\theta_i, \ell_i^1, \dots, \ell_i^{N_i})$. Evaluating the final

sequence in $\operatorname{Seq}\mathbb{R}^{N_D}$ at the last observation-time $t = L$, we get the deep LS2T map with depth-D

$$\tilde{\Phi}_{\tilde{\theta}_1,\dots,\tilde{\theta}_D} : \operatorname{Seq}\mathcal{X} \to \mathbb{R}^{n_D}.$$

Making precise how the stacking of such low-rank sequence-to-sequence transformations approximates general functions requires more tools from algebra, and we provide a rigorous quantitative statement in Appendix 3.C. Here, we just appeal to the analogy made with adding depth in neural networks mentioned earlier and empirically validate this in our experiments in Section 3.4.

3.4 Building neural networks with LS2T layers

The Seq2Tens map Φ built from a static feature map ϕ is universal if ϕ is universal, Theorem 3.2.1. NNs form a flexible class of universal feature maps with strong empirical success for data in $\mathcal{X} = \mathbb{R}^d$, and thus make a natural choice for ϕ. Combined with standard deep learning constructions, the framework of Sections 3.2 and 3.3 can build modular and expressive layers for sequence learning.

Neural LS2T layers. The simplest choice among many is to use as static feature map $\phi : \mathcal{X} = \mathbb{R}^d \to \mathbb{R}^h$ a feedforward network with depth-P, $\phi = \phi_P \circ \cdots \circ \phi_1$ where $\phi_j(\mathbf{x}) = \sigma(\mathbf{W}_j\mathbf{x} + \mathbf{b}_j)$ for $\mathbf{W}_j \in \mathbb{R}^{h\times d}$, $\mathbf{b}_j \in \mathbb{R}^h$. We can then lift this to a map $\varphi : \mathbb{R}^d \to \mathrm{T}(\mathbb{R}^h)$ as prescribed in (3.3). Hence, the resulting LS2T layer $\mathbf{x} \mapsto (\tilde{\Phi}_{\tilde{\theta}}(\mathbf{x}_1, \dots, \mathbf{x}_i))_{i=1,\dots,L}$ is a sequence-to-sequence transform $\operatorname{Seq}\mathbb{R}^d \to \operatorname{Seq}\mathbb{R}^h$ that is parametrized by $\tilde{\theta} = (\mathbf{W}_1, \mathbf{b}_1, \dots, \mathbf{W}_P, \mathbf{b}_P, \ell_1^1, \dots, \ell_1^{N_1})$.

Bidirectional LS2T layers. The transformation in (3.6) is completely causal in the sense that each step of the output sequence depends only on past information. For generative models, it can behove us to make the output depend on both past and future information, see [88, 8, 123]. Similarly to bidirectional RNNs and LSTMs [165, 90], we may achieve this by defining a bidirectional layer,

$$\tilde{\Phi}^{\mathrm{b}}_{(\tilde{\theta}_1,\tilde{\theta}_2)}(\mathbf{x}) : \operatorname{Seq}\mathbb{R}^d \to \operatorname{Seq}\mathbb{R}^{N+N'}, \quad \mathbf{x} \mapsto (\tilde{\Phi}_{\tilde{\theta}_1}(\mathbf{x}_1, \dots, \mathbf{x}_i), \tilde{\Phi}_{\tilde{\theta}_2}(\mathbf{x}_i, \dots, \mathbf{x}_L))_{i=1}^L.$$

The sequential nature is kept intact by making the distinction between what classifies as past (the first N coordinates) and future (the last N' coordinates) information. This amounts to having a form of precognition in the model, and has been applied in e.g. dynamics generation [123], machine translation [181], and speech processing [88].

Convolutions and LS2T. We motivate to replace the time-distributed feedforward layers proposed in the paragraph above by temporal convolutions (CNN) instead. Although theory only requires the preprocessing layer of the LS2T to be a static feature map, we find that it is beneficial to capture some of the sequential information in the preprocessing layer as well, e.g. using CNNs or RNNs. From a mathematical point of view, CNNs are a straightforward extension since they can be interpreted as time-distributed feedforward layers applied to the input sequence augmented with a $p \in \mathbb{N}$ number of its lags for CNN kernel size p (see [189, Appendix D.1] for further discussion).

In the following, we precede our deep LS2T blocks by one or more CNN layers. Intuitively, CNNs and LS2Ts are similar in that both transformations operate on subsequences of their input sequence. The main difference between the two lies in that *CNNs operate on contiguous subsequences*, and therefore, capture local, short-range nonlinear interactions between timesteps; *while LS2Ts (*(3.5)*) use all non-contiguous subsequences*, and hence, learn global, long-range interactions in time. This observation motivates that the inductive biases of the two types of layers (local/global time-interactions) are highly complementary in nature, and we suggest that the improvement in the experiments on the models containing vanilla CNN blocks are due to this complementarity.

3.5 Experiments

We demonstrate the modularity and flexibility of the above LS2T and its variants by applying it to

(i) multivariate time series classification,

(ii) mortality prediction in healthcare,

(iii) generative modelling of sequential data.

In all cases, we take a strong baseline model (FCN and GP-VAE, as detailed below) and upgrade it with LS2T layers. As Thm. 3.2.1 requires the Seq2Tens layers to be preceded by at least a static feature map, we expect these layers to perform best as an add-on on top of other models, which however can be quite simple, such as a CNN. The additional computation time is negligible (in fact, for FCN it allows to reduce the number of parameters significantly, while retaining performance), but it can yield substantial improvements. This is remarkable, since the original models are already state-of-the-art on well-established (frequentist and Bayesian) benchmarks.

3.5.1 Multivariate time series classification

As the first task, we consider multivariate time series classification (TSC) on an archive of benchmark datasets collected by [11]. Numerous previous publications report results on this archive, which makes it possible to compare against several well-performing competitor methods from the TSC community. These baselines are detailed in [189, Appendix E.1]. This archive was also considered in a recent popular survey paper on DL for TSC [105], from where we borrow the two best performing models as DL baselines: FCN and ResNet. The FCN is a fully convolutional network which stacks 3 convolutional layers of kernel sizes $(8, 5, 3)$ and filters $(128, 256, 128)$ followed by a global average pooling (GAP) layer, hence employing global parameter sharing. We refer to this model as FCN_{128}. The ResNet is a residual network stacking 3 FCN blocks of various widths with skip-connections in between [98] and a final GAP layer.

The FCN is an interesting model to upgrade with LS2T layers, since the LS2T also employs parameter sharing across the sequence length, and as noted previously, convolutions are only able to learn local interactions in time, that in particular makes them ill-suited to picking up on long-range autocorrelations, which is exactly where the LS2T can provide improvements. As our models, we consider three simple architectures:

(i) LS2T^3_{64} stacks 3 LS2T layers of order-2 and width-64;

(ii) FCN_{64}-LS2T^3_{64} precedes the LS2T^3_{64} block by an FCN_{64} block; a downsized version of FCN_{128};

(iii) FCN_{128}-LS2T^3_{64} uses the full FCN_{128} and follows it by a LS2T^3_{64} block as before.

Also, both FCN-LS2T models employ skip-connections from the input to the LS2T block and from the FCN to the classification layer, allowing for the LS2T to directly see the input, and for the FCN to directly affect the final prediction. These hyperparameters were only subject to hand-tuning on a subset of the datasets, and the values we considered were $H, N \in \{32, 64, 128\}$, $M \in \{2, 3, 4\}$ and $D \in \{1, 2, 3\}$, where $H, N \in \mathbb{N}$ is the FCN and LS2T width, resp., while $M \in \mathbb{N}$ is the LS2T order and $D \in \mathbb{N}$ is the LS2T depth. We also employ techniques such as time-embeddings [125], sequence differencing and batch normalization, see [189, Appendix D.1, E.1] for further details on the experiment and [189, Figure 2] in thereof for a visualization of the architectures.

Table 3.1: Posterior probabilities given by a Bayesian signed-rank test comparison of the proposed methods against the baselines. {>}, {<}, {=} refer to the respective events that the row method is better, the column method is better, or that they are equivalent.

Model	$LS2T^3_{64}$			FCN_{64}-$LS2T^3_{64}$			FCN_{128}-$LS2T^3_{64}$		
	$p(>)$	$p(=)$	$p(<)$	$p(>)$	$p(=)$	$p(<)$	$p(>)$	$p(=)$	$p(<)$
SMTS [12]	0.180	0.000	**0.820**	0.010	0.000	**0.990**	0.008	0.000	**0.992**
LPS [13]	0.191	0.002	**0.807**	0.012	0.001	**0.987**	0.006	0.001	**0.993**
mvARF [193]	0.011	0.140	**0.849**	0.000	0.126	**0.874**	0.000	0.088	**0.912**
DTW [158]	0.033	0.000	**0.967**	0.001	0.000	**0.999**	0.000	0.000	**1.000**
ARKernel [54]	0.100	0.097	**0.803**	0.000	0.021	**0.979**	0.000	0.015	**0.985**
gRSF [108]	0.481	0.011	**0.508**	0.028	0.013	**0.960**	0.022	0.013	**0.965**
MUSE [162]	0.405	0.128	**0.467**	0.001	0.074	**0.925**	0.001	0.077	**0.922**
MLSTMFCN [107]	**0.916**	0.043	0.041	0.123	0.071	**0.807**	0.055	0.110	**0.835**
FCN_{128} [200]	**0.998**	0.002	0.000	0.363	0.186	**0.451**	0.169	0.011	**0.820**
ResNet [200]	**0.998**	0.002	0.001	0.056	0.240	**0.704**	0.016	0.048	**0.935**
$LS2T^3_{64}$	-	-	-	0.000	0.001	**0.999**	0.000	0.001	**0.999**
FCN_{64}-$LS2T^3_{64}$	**0.999**	0.001	0.000	-	-	-	0.020	0.387	**0.593**

Results. We trained the models, FCN_{128}, ResNet, $LS2T^3_{64}$, FCN_{64}-$LS2T^3_{64}$, FCN_{128}-$LS2T^3_{64}$ on each of the 16 datasets 5 times while results for other methods were borrowed from the cited publications. [189, Figure 3] depicts the box-plot of distributions of accuracies and a CD diagram using the Nemenyi test [139], while [189, Table 7] shows the full list of results. Since mean-ranks based tests raise some paradoxical issues [15], it is customary to conduct pairwise comparisons using frequentist [58] or Bayesian [14] hypothesis tests. We adopted the Bayesian signed-rank test from [16], the posterior probabilities of which are displayed in Table 3.1, while the Bayesian posteriors are visualized in [189, Figure 4]. The results of the signed-rank test can be summarized as follows:

(1) $LS2T^3_{64}$ already outperforms some classic TS classifiers with high probability ($p \geq 0.8$), but it is not competitive with other DL classifiers. This observation is not surprising since even theory requires at least a static feature map to precede the LS2T.

(2) FCN_{64}-$LS2T^3_{64}$ outperforms almost all models with high probability ($p \geq 0.8$), except for ResNet (which is stil outperformed by $p \geq 0.7$), FCN_{128} and FCN_{128}-$LS2T^3_{64}$. When compared with FCN_{128}, the test is unable to decide between the two, which upon inspection of the individual results in [189, Table 7] can be explained by that on some datasets the benefit of the added LS2T block is high enough

that it outweighs the loss of flexibility incurred by reducing the width of the FCN - arguably these are the datasets where long-range autocorrelations are present in the input time series, and picking up on these improve the performance - however, on a few datasets the contrary is true.

(3) Lastly, FCN_{128}-LS2T^3_{64}, *outperforms all baseline methods with high probability* ($p \geq 0.8$), and hence successfully improves on the FCN_{128} via its added ability to learn long-range time-interactions. We remark that FCN_{64}-$LS2T^3_{64}$ *has fewer parameters than* FCN_{128} *by more than 50%*, hence we managed to compress the FCN to a fraction of its original size, while on average still slightly improving its performance, a nontrivial feat by its own accord.

3.5.2 Mortality prediction

We consider the PHYSIONET2012 challenge dataset [85] for mortality prediction, which is a case of medical TSC as the task is to predict in-hospital mortality of patients after their admission to the ICU. This is a difficult ML task due to missingness in the data, low signal-to-noise ratio (SNR), and imbalanced class distributions with a prevalence ratio of around 14%. We extend the experiments conducted in [102], which we also use as very strong baselines. Under the same experimental setting, we train two models: FCN-LS2T as ours and the FCN as another baseline. For both models, we conduct a random search for all hyperparameters with 20 samples from a pre-specified search space, and the setting with best validation performance is used for model evaluation on the test set over 5 independent model trains, exactly the same way as it was done in [102]. We preprocess the data using the same method as in [38, eq. (9)] and additionally handle static features by tiling them along the time axis and adding them as extra coordinates. We additionally introduce in both models a `SpatialDropout1D` layer after all CNN and LS2T layers with the same tunable dropout rate to mitigate the low SNR of the dataset.

Results. Table 3.2 compares the performance of FCN-LS2T with that of FCN and the results from [102] on 3 metrics:

(1) ACCURACY,

(2) area under the precision-recall curve (AUPRC),

(3) area under the ROC curve (AUROC).

Table 3.2: Comparison of FCN-LS2T and FCN on PHYSIONET2012 with the results from [102].

MODEL	ACCURACY	AUPRC	AUROC
FCN-LS2T	$\mathbf{84.1 \pm 1.6}$	$\mathbf{53.9 \pm 0.5}$	85.6 ± 0.5
FCN	80.7 ± 1.7	52.8 ± 1.3	85.6 ± 0.2
GRU-D	80.0 ± 2.9	$\mathit{53.7 \pm 0.9}$	$\mathbf{86.3 \pm 0.3}$
GRU-SIMPLE	82.2 ± 0.2	42.2 ± 0.6	80.8 ± 1.1
IP-NETS	79.4 ± 0.3	51.0 ± 0.6	$\mathit{86.0 \pm 0.2}$
PHASED-LSTM	76.8 ± 5.2	38.7 ± 1.5	79.0 ± 1.0
TRANSFORMER	$\mathit{83.7 \pm 3.5}$	52.8 ± 2.2	$\mathbf{86.3 \pm 0.8}$
LATENT-ODE	76.0 ± 0.1	50.7 ± 1.7	85.7 ± 0.6
SEFT-ATTN.	75.3 ± 3.5	52.4 ± 1.1	85.1 ± 0.4

We can observe that *FCN-LS2T takes on average first place according to both* ACCURACY *and* AUPRC*, outperforming FCN and all SOTA methods*, e.g. TRANSFORMER [196], GRU-D [38], SEFT [102], and also being competitive in terms of AUROC. This is very promising, and it suggests that LS2T layers might be particularly well-suited to complex and heterogenous datasets, such as medical time series, since the FCN-LS2T models significantly improved accuracy on ECG as well, another medical dataset in the previous experiment.

3.5.3 Generating sequential data

Finally, we demonstrate on sequential data imputation for time series and video that LS2Ts do not only provide good representations of sequences in discriminative, but also generative models.

The GP-VAE model. In this experiment, we take as base model the recent GP-VAE [73], that provides state-of-the-art results for probabilistic sequential data imputation. The GP-VAE is essentially based on the HI-VAE [138] for handling missing data in variational autoencoders (VAEs) [112] adapted to the handling of time series data by the use of a Gaussian process (GP) prior [203] across time in the latent sequence space to capture temporal dynamics. Since the GP-VAE is a highly advanced model, its in-depth description is deferred to [189, Appendix E.3]. We extend the experiments conducted in [73], and we make one simple change to the GP-VAE architecture without changing any other hyperparameters or aspects: we introduce a single bidirectional LS2T layer (B-LS2T) into the encoder network that is used in the amortized representation of the means and covariances of the variational posterior. The B-LS2T layer is preceded by a time-embedding and differencing block,

Table 3.3: Performance comparison of GP-VAE (B-LS2T) with the baseline methods

Method	HMNIST			Sprites	Physionet
	NLL	MSE	AUROC	MSE	AUROC
Mean imputation	-	0.168 ± 0.000	0.938 ± 0.000	0.013 ± 0.000	0.703 ± 0.000
Forward imputation	-	0.177 ± 0.000	0.935 ± 0.000	0.028 ± 0.000	0.710 ± 0.000
VAE	0.599 ± 0.002	0.232 ± 0.000	0.922 ± 0.000	0.028 ± 0.000	0.677 ± 0.002
HI-VAE	0.372 ± 0.008	0.134 ± 0.003	$\mathbf{0.962 \pm 0.001}$	0.007 ± 0.000	0.686 ± 0.010
GP-VAE	0.350 ± 0.007	0.114 ± 0.002	$\mathbf{0.960 \pm 0.002}$	$\mathbf{0.002 \pm 0.000}$	0.730 ± 0.006
GP-VAE (B-LS2T)	$\mathbf{0.251 \pm 0.008}$	$\mathbf{0.092 \pm 0.003}$	$\mathbf{0.962 \pm 0.001}$	$\mathbf{0.002 \pm 0.000}$	$\mathbf{0.743 \pm 0.007}$
BRITS	-	-	-	-	$\mathbf{0.742 \pm 0.008}$

and succeeded by channel flattening and layer normalization as depicted in [189, Figure 5]. The idea behind this experiment is to see if we can improve the performance of a highly complicated model that is composed of many interacting submodels, by the naive introduction of LS2T layers.

Results. To make the comparison, we ceteris paribus re-ran all experiments the authors originally included in their paper [73], which are imputation of Healing MNIST, Sprites, and Physionet 2012. The results are in Table 3.3, which report the same metrics as used in [73], i.e. negative log-likelihood (NLL, lower is better), mean squared error (MSE, lower is better) on test sets, and downstream classification performance of a linear classifier (AUROC, higher is better). For all other models beside our GP-VAE (B-LS2T), the results were borrowed from [73]. We observe that simply adding the B-LS2T layer improved the result in almost all cases, except for Sprites, where the GP-VAE already achieved a very low MSE score. Additionally, when comparing GP-VAE to BRITS on Physionet, the authors argue that although the BRITS achieves a higher AUROC score, the GP-VAE should not be disregarded as it fits a generative model to the data that enjoys the usual Bayesian benefits of predicting distributions instead of point predictions. The results display that by simply adding our layer into the architecture, we managed to elevate the performance of GP-VAE to the same level while retaining these same benefits. We believe the reason for the improvement is a tighter amortization gap in the variational approximation [51] achieved by increasing the expressiveness of the encoder by the LS2T allowing it to pick up on long-range interactions in time. We provide further discussion in [189, Appendix E.3].

3.6 Related work and Summary

Related Work. The literature on tensor models in ML is vast. Related to our approach we mention pars-pro-toto Tensor Networks [49], that use classical LR

decompositions, such as CP [36], Tucker [191], tensor trains [142] and tensor rings [206]; further, CNNs have been combined with LR tensor techniques [50, 115] and extended to RNNs [109]; Tensor Fusion Networks [205] and its LR variants [126, 124, 104]; tensor-based gait recognition [188]. Our main contribution to this literature is the use of the free algebra $\mathrm{T}(V)$ with its convolution product $\cdot$, instead of $V^{\otimes m}$ with the outer product $\otimes$ that is used in the above papers. While counter-intuitive to work in a larger space $\mathrm{T}(V)$, the additional algebra structure of $(\mathrm{T}(V), \cdot)$ is the main reason for the nice properties of Φ (*universality, making sequences of arbitrary length comparable, convergence in the continuous time limit*; see Appendix 3.B) which we believe are in turn the main reason for the *strong benchmark performance*. Stacked LR sequence transforms allow to exploit this rich algebraic structure with little computational overhead. Another related literature are path signatures in ML [130, 43, 87, 27, 190]. These arise as special case of Seq2Tens (Appendix 3.B) and our main contribution to this literature is that Seq2Tens resolves a well-known computational bottleneck in this literature since it *never needs to compute and store a signature*, instead it *directly and efficiently learns the functional of the signature*.

Summary. We used a classical non-commutative structure to construct a feature map for sequences of arbitrary length. By stacking sequence transforms we turned this into scalable and modular NN layers for sequence data. The main novelty is the use of the free algebra $\mathrm{T}(V)$ constructed from the static feature space V. While free algebras are classical in mathematics, their use in ML seems novel and underexplored. We would like to re-emphasize that $(\mathrm{T}(V), \cdot)$ is not a mysterious abstract space: if you know the outer tensor product $\otimes$ then you can easily switch to the tensor convolution product $\cdot$ by taking sums of outer tensor products, as defined in (3.2). As our experiments show, the benefits of this algebraic structure are not just theoretical but can significantly elevate performance of already strong-performing models.

For practitioners, we recommend a look at Section 3.A for a refresher on tensor notation and an introduction to $\mathrm{T}(V)$; further, the introduction of Section 3.B contains a brief summary of the main theoretical properties of Seq2Tens that make it an attractive feature map for sequence data. [189, Appendix C, D] contain details on algorithms and experiments.

For theoreticians, we recommend Section 3.B for a proof that Φ is universal (Theorem 3.B.1), how the Seq2Tens map behaves in the high-frequency limit as one goes from discrete to continuous time (Proposition 3.B.4), and to Section 3.C for a quantitative statement of low-rank functionals can be turned into high-rank functionals with sequence-to-sequence transformations. We re-emphasize that these more algebra-heavy sections are not needed for practitioners.

3.A Tensors and the Free Algebra

This section recalls some basics on the tensor product $\otimes$ and the convolution product that turns the linear space $\mathrm{T}(V)$ into an algebra – the so-called *free algebra* or *free algebra over* V. We refer to [117, Chapter 16] for more on tensors, to [152] for free algebras. Put briefly, for any linear space V there exists a linear space $\mathrm{T}(V)$ that contains V but that also carries a non-commutative product.

Tensor products on $\mathbb{R}^d$. If $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ and $y = (y_1, \dots, y_e) \in \mathbb{R}^e$ are two vectors, then their *tensor product* $x \otimes y$ is defined as the $(d \times e)$-matrix, or degree 2 tensor, with entries $(x \otimes y)_{i,j} = x_i y_j$. This is also commonly called the *outer product* of the two vectors. The space $\mathbb{R}^d \otimes \mathbb{R}^e$ is defined as the linear span of all degree 2 tensors $x \otimes y$ for $x \in \mathbb{R}^d, y \in \mathbb{R}^e$. If $z \in \mathbb{R}^f$ is another vector, then one may form a degree 3 tensor $x \otimes y \otimes z$ with shape $(d \times e \times f)$ defined to have entries $(x \otimes y \otimes z)_{i,j,k} = x_i y_j z_k$. The space $\mathbb{R}^d \otimes \mathbb{R}^e \otimes \mathbb{R}^f$ is analogously defined as the linear span of all degree 3 tensors $x \otimes y \otimes z$ for $x \in \mathbb{R}^d, y \in \mathbb{R}^e, z \in \mathbb{R}^f$.

The tensor product of two general vector spaces V and W can be defined even if they are infinite dimensional, see [117, Chapter 16], but we invite readers unfamiliar with general tensor spaces to think of V as $\mathbb{R}^d$ below.

The free algebra $\mathrm{T}(V)$. Ultimately we are not only interested in tensors of some fixed degree m – that is an element of $V^{\otimes m}$ – but sequences of tensors of increasing degree. Given some linear space V, the linear space $\mathrm{T}(V)$ is defined as set of all

tensors of any degree over V. Formally

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

where we use the notation $V^{\otimes 1} = V$, $V^{\otimes 2} = V \otimes V$, $V^{\otimes 3} = V \otimes V \otimes V$ and so on; by convention we let $V^{\otimes 0} = \mathbb{R}$. We normally write elements of $\mathrm{T}(V)$ as $\mathbf{t} = (\mathbf{t}_0, \mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3, \ldots)$ such that $\mathbf{t}_m \in V^{\otimes m}$, that is, $\mathbf{t}_0$ is a scalar, $\mathbf{t}_1$ is a vector, $\mathbf{t}_2$ is a matrix, $\mathbf{t}_3$ is a 3-tensor and so on. Note that $\mathrm{T}(V)$ is again a linear space if we define addition and scalar multiplication as

$$\mathbf{s} + \mathbf{t} = (\mathbf{s}_m + \mathbf{t}_m)_{m\geq 0} \in \mathrm{T}(V) \text{ and } c \cdot \mathbf{t} = (c\mathbf{t}_m)_{m\geq 0} \in \mathrm{T}(V)$$

for $\mathbf{s}, \mathbf{t} \in \mathrm{T}(V)$ and $c \in \mathbb{R}$.

Example 3.A.1. Let $V = \mathbb{R}^d$. For $\mathbf{v} = (\mathbf{v}_i)_{i=1,\ldots,d} \in \mathbb{R}^d$ consider $\mathbf{t} = (\mathbf{v}^{\otimes m})_{m\geq 0} \in \mathrm{T}(\mathbb{R}^d)$ where we denote for brevity

$$\mathbf{t}_m := \mathbf{v}^{\otimes m} := \underbrace{\mathbf{v} \otimes \cdots \otimes \mathbf{v}}_{m \text{ many tensor products } \otimes} \in (\mathbb{R}^d)^{\otimes m} \text{ and by convention we set } \mathbf{v}^{\otimes 0} := 1 \in (\mathbb{R}^d)^{\otimes 0}.$$

That is, $\mathbf{t}_1 = \mathbf{v}^{\otimes 1} = \mathbf{v}$ is a d-dimensional vector, with the i coordinate equal to $\mathbf{v}_i$; $\mathbf{t}_2 = \mathbf{v}^{\otimes 2}$ is $d \times d$-matrix with the (i,j)-coordinate equal to $\mathbf{v}_i\mathbf{v}_j$; $\mathbf{t}_3 = \mathbf{v}^{\otimes 3}$ is degree 3-tensor with the (i,j,k)-coordinate equal to $\mathbf{v}_i\mathbf{v}_j\mathbf{v}_k$. In this special case, the element $\mathbf{t} \in \mathrm{T}(\mathbb{R}^d)$ consists of entries $\mathbf{t}_m = \mathbf{v}^{\otimes m} \in (\mathbb{R}^d)^{\otimes m}$ that are symmetric tensors, that is the $(i_1, \ldots, i_m)$-th coordinate is the same as the $(i_{\sigma(1)}, \ldots, i_{\sigma(d)})$ coordinate if σ is a permutation of $\{1, \ldots, d\}$. However, we emphasize that in general an element of $\mathrm{T}(\mathbb{R}^d)$ does not need to be made up of symmetric tensors.

A product on $\mathrm{T}(V)$. Key to our approach is that $\mathrm{T}(V)$ is not only a linear space, but what distinguishes it as a feature space for sequences is that it carries a non-commutative product. In other words, $\mathrm{T}(V)$ is not just a vector space but a (non-commutative) algebra (an algebra is a vector space where one can multiply elements). This is the so-called *tensor convolution product* and defined as follows

$$\mathbf{s} \cdot \mathbf{t} := \Big(\sum_{i=0}^{m} \mathbf{s}_i \otimes \mathbf{t}_{m-i}\Big)_{m\geq 0} = \Big(1, \mathbf{s}_1 + \mathbf{t}_1, \mathbf{s}_2 + \mathbf{s}_1 \otimes \mathbf{t}_1 + \mathbf{t}_2, \ldots\Big).$$

In a precise mathematical sense, $\mathrm{T}(V)$ is the most general algebra containing V, namely $\mathrm{T}(V)$ *is the "free algebra" that contains* V; see [117, Chapter 16] for the precise definition of free objects.

3.B A universal feature map for sequences of arbitrary length

Recall from Section 3.2, that given a map defined on a set $\mathcal{X}$

$$\phi : \mathcal{X} \to V$$

we lift ϕ to a map $\varphi : \mathcal{X} \to \mathrm{T}(V)$ and define the Seq2Tens feature map for sequences in $\mathcal{X}$ of arbitrary length as

$$\Phi : \mathrm{Seq}\,\mathcal{X} \to \mathrm{T}(V), \quad \mathbf{x} \to \prod_{i=1}^{T} \varphi(\mathbf{x}_i).$$

The remainder of Section 3.B makes the following statements mathematically rigorous:

(i) Φ is a universal feature map whenever ϕ is a universal (Section 3.B.1 and 3.B.2),

(ii) Φ makes sequences of different length comparable analogous to how m-mers make strings of different length comparable (Section 3.B.3),

(iii) Φ converges to a well-defined object when we go from discrete to continuous time (sequences converge to paths) (Section 3.B.4).

3.B.1 The universality of Φ.

Definition 13. Let $\mathcal{X}$ be a topological space (the "data space") and W a linear space ("the feature space"). We say that a function $f : \mathcal{X} \to W$ is universal (to $C_b(\mathcal{X})$) if the set of functions

$$\{x \mapsto \langle \ell, f(x) \rangle \,:\, \ell \in W'\} \subseteq C_b(\mathcal{X})$$

is dense in $C_b(\mathcal{X})$.

Example 3.B.1. Classic examples of this in ML are

- For $\mathcal{X} \subset \mathbb{R}^d$ bounded and $W = \mathrm{T}(\mathbb{R}^d)$, the polynomial map $p : \mathbb{R}^d \to \mathrm{T}(\mathbb{R}^d), \mathbf{x} \mapsto (1, \mathbf{x}, \mathbf{x}^{\otimes 2}, \mathbf{x}^{\otimes 3}, \mathbf{x}^{\otimes 4}, \ldots)$ is universal [153].
- The 1-layer neural net map $\mathbf{x} \mapsto \prod_\theta N_\theta(\mathbf{x})$ where θ runs over all configurations of parameters is universal under some very mild conditions [103].

We now prove the main result of this section

Theorem 3.B.1. *Let $\varphi : \mathcal{X} \to \mathrm{T}(V), \mathbf{x} \mapsto (\varphi_m(\mathbf{x}))_{m\geq 0}, \varphi_m(\mathbf{x}) \in V^{\otimes m}$ be such that:*

1. *For any $n \geq 1$ the support of $(\varphi_0, \varphi_1, \ldots, \varphi_{m_1})^{\otimes n}$ and φ_{m_2} are disjoint if $1 \leq m_1 < m_2$.*
2. *$\varphi_0 = 1$ and $\varphi_1 : \mathcal{X} \to V$ is a bounded universal map with at least one constant term.*

Then

$$\Phi : \mathrm{Seq}(\mathcal{X}) \to \mathrm{T}(V), \quad (\mathbf{x}_1, \ldots, \mathbf{x}_L) \mapsto \prod_{i=1}^{L} \varphi(\mathbf{x}_i)$$

is universal.

Remark 3.B.1.

(i) Theorem3.B.1 implies that $\mathbf{x} \mapsto \prod_{i=1}^{L}(1, \phi(\mathbf{x}), 0, \ldots)$ is universal whenever $\phi : \mathcal{X} \to V$ is universal. This is the lift we use throughout the main text, see (3.3).

(ii) By taking $\varphi : \mathbb{R}^d \to \mathrm{T}(\mathbb{R}^d), \mathbf{x} \mapsto (1, \mathbf{x}, \frac{\mathbf{x}^{\otimes 2}}{2!}, \frac{\mathbf{x}^{\otimes 3}}{3!}, \ldots)$ one recovers Chen's signature [39, 40, 41] as used in rough paths.

(iii) By taking $\varphi : \mathbb{R}^d \to \mathrm{T}(V)$, $\varphi_1(\mathbf{x})$ the polynomial map and $\varphi_m(\mathbf{x}) = 0$ for $m \geq 2$ one recovers the iterated sums of [60] and [113].

(iv) By taking each φ_m to be a trainable Neural Network one gets a trainable universal map Φ for sequences that includes all of the above,

The algebra of linear functionals on Φ.

The proof of Theorem 3.B.1 uses that if φ is universal, then the space of linear functionals on $\Phi(\mathbf{x})$ forms a commutative algebra, that is for two linear functionals ℓ_1, ℓ_2 there exists another linear functional ℓ such that

$$\langle \ell_1, \Phi(\mathbf{x}) \rangle \langle \ell_2, \Phi(\mathbf{x}) \rangle = \langle \ell, \Phi(\mathbf{x}) \rangle.$$

This new functional ℓ is constructed in explicit way from ℓ_1 and ℓ_2, with a so-called quasi-shuffle product. In the remainder of this section 3.B.1, we prepare and give the proof of Theorem 3.B.1: subsection 3.B.1 introduces the quasi-shuffle product, and subsection 3.B.1 uses this to prove Theorem 3.B.1.

We spell out the proof for the case $\varphi = (1, \phi, 0, 0, \ldots) \in \mathrm{T}(V)$ since this is the form we use in the main text, Proposition 3.2.1, and the other cases follow similarly.

In fact, without loss of generality we can take $\phi = \mathrm{id}$ since this does not change the algebraic structure in any way. That is, we take

$$\Phi : \operatorname{Seq} V \to \mathrm{T}(V), \quad \Phi(\mathbf{x}_1, \dots, \mathbf{x}_L) := \prod_{i=1}^{L} \varphi(\mathbf{x}_i) \tag{3.7}$$

with $\varphi(\mathbf{x}) = (1, \mathbf{x}, 0, 0, \dots)$. By using the definition of the product in $\mathrm{T}(V)$ and expanding (3.7) we get

$$\Phi(\mathbf{x}_1, \dots, \mathbf{x}_L) = (1, \sum_{i=1}^{L} \underbrace{\mathbf{x}_i}_{V}, \sum_{1 \le i_1 < i_2 \le L} \underbrace{\mathbf{x}_{i_1} \otimes \mathbf{x}_{i_2}}_{V^{\otimes 2}}, \sum_{1 \le i_1 < i_2 < i_3 \le L} \underbrace{\mathbf{x}_{i_1} \otimes \mathbf{x}_{i_2} \otimes \mathbf{x}_{i_3}}_{V^{\otimes 3}}, \cdots)$$

In general, writing $\Phi_m(\mathbf{x})$ for the projection of $\Phi(\mathbf{x})$ onto $V^{\otimes m}$, we have

$$\Phi_m(\mathbf{x}) = \sum_{1 \le i_1 < \cdots < i_m \le L} \mathbf{x}_{i_1} \otimes \cdots \otimes \mathbf{x}_{i_m}. \tag{3.8}$$

So if $\ell = (0, 0, \dots, \mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_m, 0, \dots)$ with $\mathbf{v}_1, \dots, \mathbf{v}_m \in V$, then

$$\begin{aligned} \langle \ell, \Phi_m(\mathbf{x}) \rangle &= \langle \mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_m, \sum_{1 \le i_1 < \cdots < i_m \le L} \mathbf{x}_{i_1} \otimes \cdots \otimes \mathbf{x}_{i_m} \rangle \\ &= \sum_{1 \le i_1 < \cdots < i_m \le L} \langle \mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_m, \mathbf{x}_{i_1} \otimes \cdots \otimes \mathbf{x}_{i_m} \rangle = \sum_{1 \le i_1 < \cdots < i_m \le L} \langle \mathbf{v}_1, \mathbf{x}_{i_1} \rangle \cdots \langle \mathbf{v}_m, \mathbf{x}_{i_m} \rangle. \end{aligned}$$

Hence $\langle \ell, \Phi_m(\mathbf{x}) \rangle$ can be computed efficiently without computing $\Phi(\mathbf{x})$. Proposition 3.3.1 follows by linearity since by definition $\langle \ell, \Phi(\mathbf{x}) \rangle = \sum_{m \ge 0} \langle \ell_m, \Phi(\mathbf{x}) \rangle$ and for each of the terms we can use the above formula when $\ell = (\ell_0, \ell_1, \ell_2, \dots, \ell_M, 0, \dots)$ is of rank-1 and of degree M (Definition 12).

Non-linear functionals acting on Φ. We now investigate what happens when one applies non-linear functions to $\Phi(\mathbf{x})$. To do this, we first note that since $\mathrm{T}(V)$ is a vector space, we may form the free algebra over $\mathrm{T}(V)$, denoted by $\mathrm{T}(\mathrm{T}(V))$, or $\mathrm{T}^2(V)$. It may be decomposed as

$$\mathrm{T}^2(V) = \prod_{n_1, \dots, n_k \ge 0} V^{\otimes n_1} \Big| \cdots \Big| V^{\otimes n_k}$$

where we use the notation $\otimes$ for the tensor product on V and the bar $|$ for the tensor product on $\mathrm{T}(V)$. See [64] for more on $\mathrm{T}^2(V)$ and the bar notation.

Definition 14. If $x \in V$ is a vector, we denote by $x^\star$ its extension

$$x^\star := (x^{\otimes m})_{m \ge 0} = (1, x, x^{\otimes 2}, x^{\otimes 3}, \dots)$$

and if $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_L) \in \operatorname{Seq} V$ is a sequence, then

$$\mathbf{x}^\star := (\mathbf{x}_1^\star, \dots, \mathbf{x}_L^\star) \in \operatorname{Seq} \mathrm{T}(V)$$

Since $\mathbf{x}^\star$ is a sequence in $\mathrm{T}(V)$, we may compute $\Phi(\mathbf{x}^\star)$ which takes values in $\mathrm{T}(\mathrm{T}(V)) = \mathrm{T}^2(V)$. The reason for the above definition is that when products of linear functions in $\mathrm{T}(V)$ act on $\Phi(\mathbf{x})$, they may be described as linear functions in $\mathrm{T}^2(V)$ acting on $\Phi(\mathbf{x}^\star)$. That is, $\mathrm{T}(V)$ is not big enough to capture all non-linear functions acting on $\Phi(\mathbf{x})$, but $\mathrm{T}^2(V)$ is.

Definition 15. Assume that V has basis $e_1, \ldots, e_d$. The *quasi-shuffle* product

$$\star : \mathrm{T}^2(V) \times \mathrm{T}^2(V) \to \mathrm{T}^2(V)$$

is defined inductively on rank 1 elements $\ell_1 = e_{i_1}|\cdots|e_{i_m}, \ell_2 = e_{j_1}|\cdots|e_{j_n}$ by

$$(\ell_1|e_i) \star (\ell_2|e_j) = (\ell_1|e_i \star \ell_2)|e_j + (\ell_1 \star \ell_2|e_j)|e_i + (\ell_1 \star \ell_2)|(e_i \otimes e_j).$$

By linearity $\star$ extends to a product on all of $\mathrm{T}(V)$.

Lemma 3.B.2. *The map Φ satisfies the following*

$$\langle \ell_1, \Phi(\mathbf{x})\rangle\langle \ell_2, \Phi(\mathbf{x})\rangle = \langle \ell_1 \star \ell_2, \Phi(\mathbf{x}^\star)\rangle.$$

Proof. By writing out (3.8) in coordinates we get

$$\langle e_{i_1}|\cdots|e_{i_m}, \Phi(\mathbf{x})\rangle = \sum_{1 \le k_1 < \cdots < k_m \le L} \langle e_{i_1}, \mathbf{x}_{k_1}\rangle \cdots \langle e_{i_m}, \mathbf{x}_{k_m}\rangle.$$

which shows that Φ satisfies a recurrence equation. The proof follows by induction. □

The space $\mathrm{T}^2(V)$ might seem very large and difficult to work with at first. The power of this representation comes from the fact that one may leverage this in proving strong statements about the original map $\Phi : \mathrm{Seq}\, V \to \mathrm{T}(V)$, and we will use this in the next subsection.

3.B.2 Proof of Theorem 3.B.1.

We prepare the proof of Theorem 3.B.1 with the following lemma.

Lemma 3.B.3. *Let $\mathrm{Seq}^1(V)$ be the set of all $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_L) \in \mathrm{Seq}\, V$ with the form $x_i = (1, x_i^1, \ldots, x_i^d)$. That is, all sequences where one of the terms is constant. Then the map*

$$\mathrm{Seq}^1(V) \to \mathrm{T}(V), \quad (\mathbf{x}_1, \ldots, \mathbf{x}_L) \to \prod_{i=1}^{L} (1 + \mathbf{x}_i)$$

is injective.

Proof. Follows from an induction argument over L. For $L = 1$ it is clear since

$$\langle e_i, \Phi(\mathbf{x})\rangle = x_i.$$

Assume that it is true for L, let $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_{L+1}), \mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_{L+1})$, where we may assume that both have length $L+1$ by taking any number of components to be 0 if necessary. Let ℓ_1 be some linear function that separates $\Phi(\mathbf{x}_1, \ldots, \mathbf{x}_L)$ and $\Phi(\mathbf{y}_1, \ldots, \mathbf{y}_L)$ and ℓ_2 some linear function that separates $\Phi(\mathbf{x}_2, \ldots, \mathbf{x}_{L+1})$ and $\Phi(\mathbf{y}_2, \ldots, \mathbf{y}_{L+1})$, then by fixing some $\gamma \in \mathbb{R}$:

$$\begin{aligned}&\langle \ell_1 \otimes e_0 + \gamma e_0 \otimes \ell_2, \Phi(\mathbf{x}) - \Phi(\mathbf{y})\rangle \\ &= \langle \ell_1, \Phi(\mathbf{x}_1, \ldots, \mathbf{x}_L) - \Phi(\mathbf{y}_1, \ldots, \mathbf{y}_L)\rangle + \gamma\langle \ell_2, \Phi(\mathbf{x}_2, \ldots, \mathbf{x}_{L+1}) - \Phi(\mathbf{y}_2, \ldots, \mathbf{y}_{L+1})\rangle.\end{aligned}$$

Since neither $\langle \ell_1, \Phi(\mathbf{x}_1, \ldots, \mathbf{x}_L) - \Phi(\mathbf{y}_1, \ldots, \mathbf{y}_L)\rangle$ nor $\langle \ell_2, \Phi(\mathbf{x}_2, \ldots, \mathbf{x}_{L+1}) - \Phi(\mathbf{y}_2, \ldots, \mathbf{y}_{L+1})\rangle$ are 0 by assumption there exists some $\gamma \in \mathbb{R}$ such that $\langle \ell_1 \otimes e_0 + \gamma e_0 \otimes \ell_2, \Phi(\mathbf{x}) - \Phi(\mathbf{y})\rangle \neq 0$. This shows the assertion. □

We now have everything to give a proof of Theorem 3.B.1.

Proof of Theorem 3.B.1. We will show that linear functionals on Φ are dense in the *strict topology* [82]. By Theorem [82, Theorem 3.1] it is enough to show that linear functions on Φ form an algebra since by Lemma 3.B.3 they separates the points of $\operatorname{Seq} \mathcal{X}$. Since they clearly form a vector space it is enough to show that they are closed under point-wise multiplication. Let ℓ_1, ℓ_2 be two such, then by Lemma 3.B.2

$$\langle \ell_1, \Phi(\mathbf{x})\rangle\langle \ell_2, \Phi(\mathbf{x})\rangle = \langle \ell_1 \star \ell_2, \prod_{i=1}^{L} \phi(\mathbf{x}_i)^\star\rangle$$

so it is enough to show that $\ell_1 \star \ell_2$ also is a linear function on $\Phi(\mathbf{x})$. Note that inductively it is enough to show that if e_i, e_j are unit vectors, then $e_i \otimes e_j$ is a linear function on $\Phi(\mathbf{x})$. By assumption ϕ is bounded and universal, so the continuous bounded function $\mathbf{x} \mapsto \langle e_i, \phi(\mathbf{x}_k)\rangle\langle e_j, \phi(\mathbf{x}_k)\rangle$ is approximately linear, and we may write

$$\langle e_i, \phi(\mathbf{x}_k)\rangle\langle e_j, \phi(\mathbf{x}_k)\rangle = \langle h, \phi(\mathbf{x}_k)\rangle + \varepsilon(\mathbf{x}_k)$$

where $\varepsilon(\mathbf{x}_k)$ can be made arbitrarily small in the strict topology. The assertion now follows since

$$\begin{aligned}&\langle e_i \otimes e_j, \prod_{i=1}^{L} \phi(\mathbf{x}_i)^\star\rangle = \sum_{k=1}^{L} \langle e_i, \phi(\mathbf{x}_k)\rangle\langle e_j, \phi(\mathbf{x}_k)\rangle = \sum_{k=1}^{L} \langle h, \phi(\mathbf{x}_k)\rangle + \varepsilon(\mathbf{x}_k) \\ &= \langle e_h, \Phi(\mathbf{x})\rangle + n \max_{1 \leq k \leq n} \varepsilon(\mathbf{x}_k).\end{aligned}$$

□

3.B.3 Seq2Tens makes sequences of different length comparable

The simplest kind of a sequence is a string, that is a sequence of letters. Strings are determined by

(i) what letters appear in them,

(ii) in what order the letters appear.

A classical way to produce a graded description of strings is by counting their non-contiguous sub-strings. These are the so-called *m-mers*; for example,

The string "aabc" has the 2-mers $\{aa, ab, ac, ab, ac, bc\}$.

Measuring similarity between strings by counting how many substrings they have in common is a sensible similarity measure, even if the strings have different length; we refer to [122] for applications and to [52, Chapter 11] for detailed introduction to the use of substrings in ML.

Our Seq2Tens feature map can be regarded as a vast generalization of such subpattern matching: if $\Phi(\mathbf{x}) = (\Phi_m(\mathbf{x}))_{m\geq 0} \in \mathrm{T}(V)$ then the tensor $\Phi_m(\mathbf{x}) \in V^{\otimes m}$ represents non-contiguous sub-sequences of $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_T)$ of length m and thus comparing $\Phi_m(\mathbf{x})$ and $\Phi_m(\mathbf{y})$ is meaningful even when $\mathbf{x}$ and $\mathbf{y}$ are of different length. It is instructive to spell out in detail how m-mers are a special case of Seq2Tens, Example 3.B.2, and how it generalizes, Example 3.B.3.

Example 3.B.2. Let $\mathcal{X} = \{a, b, c\}$ and $\phi : \mathcal{X} \to V = \mathbb{R}^3$ defined by mapping $a, b, c \in \mathcal{X}$ to the unit vectors $e_1, e_2, e_3 \in V$, so that $\varphi(a) = (1, e_1, 0, 0, ...) \in \mathrm{T}(V)$, $\varphi(b) = (1, e_2, 0, \dots) \in \mathrm{T}(V)$, and $\varphi(c) = (1, e_3, 0, \dots) \in \mathrm{T}(V)$. For the sequence $\mathbf{x} = (a, a, b, c) \in \mathrm{Seq}\,\mathcal{X}$ we get

$$\begin{aligned}
\Phi(\mathbf{x}) =& \varphi(a)\varphi(a)\varphi(b)\varphi(c) \\
=& (1, e_1, 0, 0, \dots,) \cdot (1, e_1, 0, 0, \dots) \cdot (1, e_2, 0, 0, \dots) \cdot (1, e_3, 0, 0, \dots) \\
=& (1, \underbrace{2e_1 + e_2 + e_3}_{\in V}, \underbrace{e_1 \otimes e_1 + 2e_1 \otimes e_2 + 2e_1 \otimes e_3 + e_2 \otimes e_3}_{\in V^{\otimes 2}}, \\
& \underbrace{e_1 \otimes e_1 \otimes e_2 + e_1 \otimes e_1 \otimes e_3, e_1 \otimes e_1 \otimes e_3 \otimes e_4}_{\in V^{\otimes 3}}, \underbrace{e_1 \otimes e_1 \otimes e_2 \otimes e_3}_{\in V^{\otimes 4}}, 0, \dots).
\end{aligned}$$

We see that the tensor $\Phi(\mathbf{x}) \in V^{\otimes m}$ of degree m contains the m-mers, that is the coordinates of $\Phi_m(\mathbf{x})$ count how often a subsequence of length m in $\mathbf{x} = (a, a, b, c)$ appears; e.g. the coordinate $e_1 \otimes e_2$ of $\Phi_2(\mathbf{x})$ equals 2 because the substring "a,b"

appears twice but the $e_1 \otimes e_1$ coordinate of $\Phi_2(\mathbf{x})$ equals 1 since the substring "a,a" appears only once in (a, a, b, c), etc. Similarly, the only non-zero coordinate of $\Phi_4(\mathbf{x})$ is $e_1 \otimes e_1 \otimes e_2 \otimes e_3$ since "a,a,b,c" is the only substring of (a, a, b, c) and consequently all coordinate of $\Phi_m(\mathbf{x})$ are 0 for $m \geq 5$.

An analogous calculation shows that for continuous domain such as $\mathcal{X} = \mathbb{R}^d$ the coordinates of $\Phi_m(\mathbf{x}) \in (\mathbb{R}^d)^{\otimes m}$ measure movement patterns in these coordinates. Section 3.B.4 below shows that this interpretation also holds when we go from discrete time (sequences) to continuous time (paths); Example 3.B.3.

3.B.4 Convergence from discrete to continuous time

A common source of of sequence data is to measure a quantity $(x(t))_{t\in[0,T]}$ that evolves in continuous time at fixed times $t_1, \ldots, t_L$ to produce a sequence $\mathbf{x} = (x(t_1), \ldots, x(t_L)) \in \operatorname{Seq} \mathcal{X}$. Often the measurements are of high-frequency ($|t_{i+1} - t_i| \to 0$ and $L \to \infty$) and is interesting to understand how our Seq2Tens approach behaves in this limiting case. As it turns out, when combined with taking finite differences[1] our feature map $\Phi(\mathbf{x})$ converges to a classical object in analysis, the so-called *signature of the path* x in this limit, [43] . For brevity, we spell it out here for smooth paths and with the lift $\varphi(x) = (1, \phi(x), 0, \ldots)$, but readers familiar with rough paths will notice that the result generalizes even to non-smooth paths such as Brownian motion.

Proposition 3.B.4. *Let $x \in C^1([0, 1], \mathbb{R}^d)$ and for every $L \geq 1$ define*

$$\mathbf{x}^L := (x(t_0^L), x(t_1^L) - x(t_0^L), \ldots, x(t_L^L) - x(t_{L-1}^L))$$

where $t_i := \frac{i}{L}$ for $i = 0, \ldots, L$. Then for every $m \geq 0$ we have

$$\Phi_m(\mathbf{x}^L) \to \int_{0\leq t_1 \leq \cdots \leq t_m \leq 1} \frac{dx}{dt}(t_1) \otimes \cdots \otimes \frac{dx}{dt}(t_m) dt_1 \cdots dt_m \text{ as } L \to \infty.$$

Proof. Using the recurrence relation from the proof of Lemma 3.B.2 we may write

$$\Phi_m(\mathbf{x}^L) = \sum_{i=1}^{L} \Phi_{m-1}(\mathbf{x}_i^L) \otimes \Big(x(t_{i+1}) - x(t_i)\Big)$$

where $\mathbf{x}_i^L$ denotes the sequence $(x(t_0^L), x(t_1^L) - x(t_0^L), \ldots, x(t_i^L) - x(t_{i-1}^L))$. By a Taylor expansion this is equal to

$$\Phi_m(\mathbf{x}^L) = \sum_{i=1}^{L} \Phi_{m-1}(\mathbf{x}_i^L) \otimes \frac{dx}{dt}(t_i)(t_{i+1} - t_i) + O(1/L),$$

[1] This is necessary to counteract the fact the magnitude of Φ grows with the sum of the elements in the sequence

so by unravelling the recurrence relation we may write

$$\Phi_m(\mathbf{x}^L) = \sum_{1\le i_1<\dots<i_m\le L} \frac{dx}{dt}(t_{i_1}) \otimes \cdots \otimes \frac{dx}{dt}(t_{i_m})(t_{i_1+1} - t_{i_1}) \cdots (t_{i_m+1} - t_{i_m}) + O(1/L),$$

which is a Riemann sum, and thus converges to the asserted limit. □

The interpretation of $\Phi_m(\mathbf{x})$ as counting sub-patterns remains even in the continuous time case:

Example 3.B.3. Let $x \in C^1([0,1], \mathbb{R}^d)$ and $e_1, \dots, e_d$ the standard basis of $V = \mathbb{R}^d$. For $m = 1$ the e_1 coordinate equals

$$\langle e_1, \int_{t=0}^{1} \frac{dx}{dt}(t)dt \rangle = \langle e_1, x(1) - x(0) \rangle$$

which measures the total movement of the path x in the direction e_1. Analogously, for $m = 2$ the $e_1 \otimes e_2$ coordinate equals

$$\langle e_1 \otimes e_2, \int_{0\le t_1\le\cdots\le t_m\le 1} \frac{x(t_1)}{dt_1} \otimes \frac{x(t_2)}{dt_2} dt_1 dt_2 = \int_{0\le t_1\le t_2\le 1} \langle e_1, \frac{x(t_1)}{dt_1} \rangle \langle e_2, \frac{x(t_2)}{dt_2} \rangle dt_1 dt_2.$$

which measure the number of ordered tuples (t_1, t_2), $t_1 < t_2$, such that x moves in direction e_1 at time t_1 and subsequently in direction e_2 at time t_2.

3.C Stacking sequence-to-sequence transforms

As Φ applies to sequences of any length, we may use it to map the original sequence to another sequence in feature space,

$$\begin{aligned} \operatorname{Seq} V &\to \operatorname{Seq} \mathrm{T}(V) \\ (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_L) &\mapsto \Big(\Phi(\mathbf{x}_1), \Phi(\mathbf{x}_1, \mathbf{x}_2), \Phi(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3), \dots, \Phi(\mathbf{x}_1, \dots, \mathbf{x}_L) \Big). \end{aligned}$$

Since $\mathrm{T}(V)$ is again a linear space, we can repeat this procedure to map $\operatorname{Seq} \mathrm{T}(V)$ to $\operatorname{Seq} \mathrm{T}(\mathrm{T}(V))$. By repeating this D times, we have constructed sequence-to-sequence transforms

$$\operatorname{Seq} V = \operatorname{Seq} \mathrm{T}(V) \to \operatorname{Seq} \mathrm{T}(\mathrm{T}(V)) \to \cdots \to \operatorname{Seq} \underbrace{T(\cdots T}_{D \text{ times}}(V) \cdots).$$

See Figure 3.1 for an illustration. We emphasize that in each step of the iteration, the newly created sequence evolves in a much richer space than in the previous step. To make this precise we now introduce the *higher rank free algebras*.

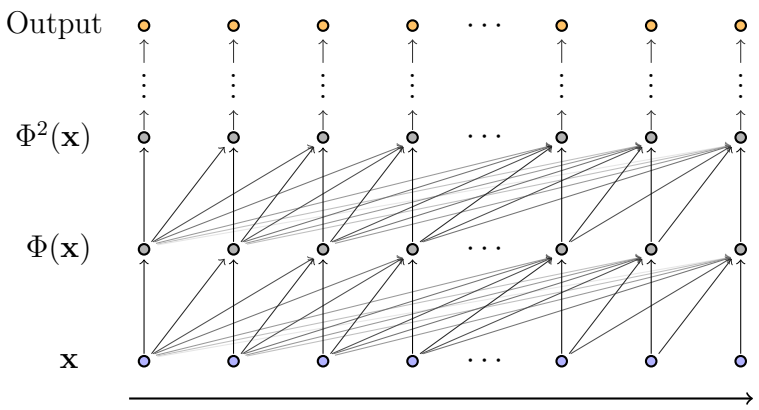


Figure 3.1: The sequence-to-sequence transformation. The bottom row is the original sequence $\mathbf{x}$ and subsequent ones apply the map Φ in the sequence-to-sequence manner.

Higher rank free algebras. Just like in Appendix 3.B we need to enlarge the ambient space $\mathrm{T}(V)$. Recall that we defined $\mathrm{T}^2(V) := \mathrm{T}(\mathrm{T}(V))$. This construction can be iterated indefinitely and leads to the higher rank free algebras, recursively defined as follows:

Definition 16. Define the spaces

$$\mathrm{T}^0(V) = V, \quad \mathrm{T}^D(V) = \prod_{m \geq 0} \left(\mathrm{T}^{D-1}(V)\right)^{\otimes m}.$$

We use the notation $\otimes_{(D)}$ for the tensor product on $\mathrm{T}^{D-1}(V)$ which makes $\left(\mathrm{T}^D(V), +, \otimes_{(D)}\right)$ into a multi-graded algebra over V. See [28] for more on this iterated construction.

Half-shuffles. By iterating the sequence-to-sequence D times, one gets a map

$$\Phi^D : \mathrm{Seq}\, V \to \mathrm{T}^D(V).$$

These are very large spaces, but as we will see in Proposition 3.C.1 below, linear functionals on the full map $\mathrm{Seq}\, V \to \mathrm{T}^D(V)$ can be de-constructed into so called *half-quasi-shuffle* on the original map $\Phi : \mathrm{Seq}\, V \to \mathrm{T}(V)$.

Just like in Appendix 3.B we consider the sequence $\mathbf{x}$ as its extension $\mathbf{x}^\star$ taking values in $\mathrm{T}(V)$. Hence linear functionals can be written as linear combinations of elements of the form $e_{i_1} \otimes_{(2)} \cdots \otimes_{(2)} e_{i_n}$.

Definition 17. The half-quasi-shuffle product is defined on rank 1 tensors by

$$\ell_1 \prec (\ell_2 \otimes_{(2)} e_i) = (\ell_1 \star \ell_2) \otimes_{(2)} e_i$$

and extends by bi-linearity to a map $\mathrm{T}^2(V) \times \mathrm{T}^2(V) \to \mathrm{T}^2(V)$.

Proposition 3.C.1 shows that by composing Φ with itself, low degree tensors on the second level can be rewritten as higher degree tensors on the first level. This indicates that iterated compositions of Φ can be much more efficient than computing everything on the first level. We show this for the first level, but by iterating the statement it can be applied for any number $D \geq 2$.

Proposition 3.C.1. *Let be the sequence-to-sequence transformation:*

$$\operatorname{Seq} V \to \operatorname{Seq} \mathrm{T}(V), \quad (\mathbf{x}_1, \dots, \mathbf{x}_L) \mapsto (\Phi(\mathbf{x}_1), \dots, \Phi(\mathbf{x}, \dots, \mathbf{x}_L))$$

and let $\Delta(\mathbf{x}_0, \dots, \mathbf{x}_L) = (\mathbf{x}_1 - \mathbf{x}_0, \dots, \mathbf{x}_L - \mathbf{x}_{L-1})$. *Then*

$$\langle e_{\ell_1} \otimes_{(3)} e_{\ell_2}, \Phi(\Delta(\mathbf{x}))\rangle = \langle \ell_1 \prec \ell_2, \Phi(\mathbf{x})\rangle$$

Proof. We use the notation $\Phi(\mathbf{x})_k = \Phi(\mathbf{x}_1, \dots, \mathbf{x}_k)$. By induction:

$$\begin{aligned}
&\langle e_{\ell_1} \otimes_{(3)} e_{\ell_2 \otimes_{(2)} e_i}, \Phi(\Delta(\mathbf{x}))\rangle \\
&= \sum_{1 \leq k_1 < k_2 \leq L} \Big(\langle \ell_1, \Phi(\mathbf{x})_{k_1}\rangle - \langle \ell_1, \Phi(\mathbf{x})_{k_1 - 1}\Big)\rangle \Big(\langle \ell_2 \otimes_{(2)} e_i, \Phi(\mathbf{x})_{k_2}\rangle - \langle \ell_2 \otimes_{(2)} e_i, \Phi(\mathbf{x})_{k_2 - 1}\rangle\Big) \\
&= \sum_{k=1}^{L-1} \langle \ell_1, \Phi(\mathbf{x})_k\rangle \Big(\langle \ell_2 \otimes_{(2)} e_i, \Phi(x)_{k+1}\rangle - \langle \ell_2 \otimes_{(2)} e_i, \Phi(\mathbf{x})_k\rangle\Big) \\
&= \sum_{k=1}^{L-1} \langle \ell_1, \Phi(\mathbf{x})_k\rangle \Big(\sum_{1 \leq l \leq k} \langle \ell_2, \Phi(\mathbf{x})_l\rangle \mathbf{x}^i_{l+1} - \langle \ell_2, \Phi(\mathbf{x})_{l-1}\rangle \mathbf{x}^i_l\Big) \\
&= \sum_{k=1}^{L-1} \langle \ell_1, \Phi(\mathbf{x})_k\rangle \langle \ell_2, \Phi(\mathbf{x})_k\rangle \mathbf{x}^i_{k+1} = \sum_{k=1}^{L-1} \langle \ell_1 \star \ell_2, \Phi(\mathbf{x})_k\rangle \mathbf{x}^i_{k+1} \\
&= \langle (\ell_1 \star \ell_2) \otimes_{(2)} e_i, \Phi(\mathbf{x})_L\rangle.
\end{aligned}$$

□

Chapter 4

Adapted topologies and higher rank signatures

This chapter is based on [28] which was written with Harald Oberhauser and Chong Liu.

Two adapted stochastic processes can have similar laws but give different results in applications such as optimal stopping, queuing theory, or stochastic programming. The reason is that the topology of weak convergence does not account for the growth of information over time that is captured in the filtration of an adapted stochastic process. To address such discontinuities, Aldous introduced the extended weak topology, and subsequently, Hoover and Keisler showed that both, weak topology and extended weak topology, are just the first two topologies in a sequence of topologies that get increasingly finer. We introduce higher rank expected signatures to embed adapted processes into graded linear spaces and show that these embeddings induce the adapted topologies of Hoover–Keisler.

4.1 Introduction

A sequence of $\mathbb{R}^d$-valued random variables $(X_n)_{n\geq 0}$ is said to converge weakly to a random variable X if

$$\int_{\mathbb{R}^d} f(x)\mathbb{P}_n(X_n \in dx) \to \int_{\mathbb{R}^d} f(x)\mathbb{P}(X \in dx) \quad \text{for all } f \in C_b(\mathbb{R}^d, \mathbb{R}).$$

If one replaces $\mathbb{R}^d$-valued random variables by path-valued random variables – that is random maps from a totally ordered set I into $\mathbb{R}^d$ – one arrives at the definition of weak convergence of stochastic processes. However, reducing stochastic processes to path-valued random variables ignores the filtration of the process. Filtrations encode how the information one has observed in the past restricts the future possibilities and thereby encodes actionable information. Even for discrete-time and real-valued Markov processes equipped with their natural filtrations, the weak topology is sometimes too coarse, see Example 4.1.1.

Example 4.1.1 ([3, 6])**.**

1. The value map of an optimal stopping problem,

$$\mathbf{X} := (\Omega, (\mathcal{F}_t)_{t\in I}, \mathbb{P}, (X_t)_{t\in I}) \mapsto \inf_{\gamma} \mathbb{E}[L_\gamma], \tag{4.1}$$

   where the inf is taken over all stopping times $\gamma \leq T$ is not continuous in the weak topology if L is an adapted functional that depends continuously on the sample path of $\mathbf{X}$. This discontinuity remains even if the domain of the solution map (4.1) is restricted to the space of discrete-time Markov processes equipped with their natural filtration.

2. Figure 4.1 shows a sequence of Markov processes that converge weakly. However, at time $t = 1$ one would make very different decisions upon observing the process for finite n and its weak limit (e.g. for portfolio allocations of investments or in optimal stopping problems such as (4.1)). The reason for this discontinuity is that although the law of the processes gets arbitrarily close for large n, their natural filtrations are very different.

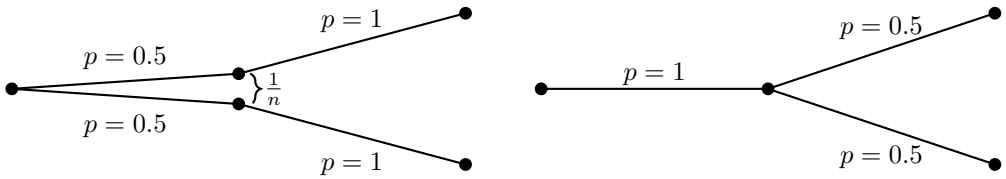


Figure 4.1: A typical example of when weak convergence is insufficient. The process on the left can be made arbitrarily close to the process on the right as $n \to \infty$.

4.1.1 Adapted Topologies

Such shortcomings of weak convergence for stochastic processes were recognized and addressed in the 1970's and 1980's. Denote by $\mathcal{S}$ the set of adapted processes that evolve in a state space that is compact subset of $\mathbb{R}^d$. David Aldous proposed to associate with an adapted process $\mathbf{X} = (\Omega, (\mathcal{F}_t)_{t\in I}, \mathbb{P}, (X_t)_{t\in I}) \in \mathcal{S}$ its *prediction process* $\hat{\mathbf{X}} = (\Omega, (\mathcal{F}_t)_{t\in I}, \mathbb{P}, (\hat{X}_t)_{t\in I})$,

$$\hat{X}_t := \mathbb{P}(X \in \cdot | \mathcal{F}_t),$$

and to define a topology on $\mathcal{S}$ by prescribing that two processes converge if and only if their prediction processes converge in the weak topology (that is, weak convergence in the space of measure-valued processes). Aldous studied this topology in [3] and showed that it has several attractive properties such as making the map in Example 4.1.1 item 1 continuous and separating the two processes in item 2. Similar points were also made and further developed by a number of different researchers [198, 199, 118, 154, 197, 66, 7] including ones in other communities such as economics [100], operations research [146, 150, 147, 148, 149], and numerics [21] and has led to the development of topologies that are finer than the classical weak topology. The construction of all these differ in detail, but in discrete time and under the natural filtration they lead to the same topology that Aldous originally introduced as was recently shown in [6]. We henceforth refer to this topology[1] as the *adapted topology of rank* 1 and we refer to the classic weak topology as the *adapted topology of rank* 0 (denoted τ_1 and τ_0 respectively).

However, even the adapted topology of rank 1 (weak convergence of the prediction process) does not characterize the full structure of adapted processes, as evidenced by Example 4.1.2

Example 4.1.2 (Example 3.2, [101])**.** There exists two sequences of Markov chains, $(\mathbf{X}_n)_n$ and $(\mathbf{Y}_n)_n$, that both converge to the same process in the topology τ_0 and in

[1] Aldous refers to this topology as the *extended weak topology*.

the topology τ_1 as $n \to \infty$. However, the information contained in their filtrations is still different; for example $\mathbb{E}[\mathbb{E}[\mathbf{X}_4^n|\mathcal{F}_3]^2|\mathcal{F}_1] - \mathbb{E}[\mathbb{E}[\mathbf{Y}_4^n|\mathcal{F}_3]^2|\mathcal{F}_1] \not\to 0$ as $n \to \infty$; see Appendix 4.A for details.

Seminal work of [101] provides a definite answer: it shows the existence of a sequence of topologies $(\tau_r)_{r\geq 0}$ on $\mathcal{S}$ that become strictly finer as r increases; τ_0 is the topology of weak convergence; τ_1 is Aldous' weak convergence of prediction processes, and $\bigcap_{r=0}^{\infty} \tau_r$ identifies two process if and only if they are isomorphic, see [101] for the precise statement. We refer to τ_r as the *adapted topology of rank* r on $\mathcal{S}$. The approach of [101] is different than Aldous' approach that relies on prediction processes. The starting point of [101] is that one may specify a topology by choosing a class of functionals on pathspace, that is a subset of $(\mathbb{R}^d)^I \to \mathbb{R}$, and define convergence of a sequence of processes $\mathbf{X}_n$ to $\mathbf{X}$ by requiring

$$\mathbf{X}_n \to \mathbf{X}, \text{ if and only if } \int_{(\mathbb{R}^d)^I} f(x)\mathbb{P}_n(X_n \in dx) \to \int_{(\mathbb{R}^d)^I} f(x)\mathbb{P}(X \in dx),$$

for all f in this set of functionals. By taking this set of functionals to be $C_b((\mathbb{R}^d)^I, \mathbb{R})$ one recovers weak convergence, but much richer classes of functionals can be constructed by iterating conditional expectations and compositions with bounded continuous functions, e.g. $f(\mathbf{X})(\omega) = \mathbb{E}[\cos(X_{t_1}X_{t_2})|\mathcal{F}_{t_3}](\omega)$ is one such function. In fact, to avoid measure-theoretic trouble, it is more convenient to work with random variables: one defines so-called adapted functionals AF as maps from $\mathcal{S}$ to the space of real-valued random variables, by mapping a process $\mathbf{X}$ to a random variable given as above by iteration of finite marginals, conditional expectations, and continuous functions. The minimal number of nested conditional expectations needed to specify an element of AF induces the natural grading

$$\mathrm{AF} = \bigcup_{r\geq 0} \mathrm{AF}_r$$

where AF_r denotes all adapted functionals that are build with r nested conditional expectations. Hoover–Keisler showed that by defining[2]

$$\mathbf{X}_n \to \mathbf{X} \text{ if and only if } \mathbb{E}[f(\mathbf{X}_n)] \to \mathbb{E}[f(\mathbf{X})] \text{ for all } f \in \mathrm{AF}_r,$$

then the associated topologies get strictly finer as $r \to \infty$; e.g. the topology τ_2 separates the two adapted processes in Example 4.1.2.

[2]Rigorously speaking we should write it as $\mathbb{E}_{\mathbb{P}_n}[f(\mathbf{X}_n)]$ for $\mathbf{X}_n = (\Omega_n, (\mathcal{F}_t^n)_{t\in I}, \mathbb{P}_n, X_n)$. For simplicity we will omit the subscript $\mathbb{P}_n$ throughout this paper.

4.1.2 Contribution

Denote by $\mathcal{S} = \mathcal{S}(\mathbb{R}^d)$ the set of adapted stochastic processes $\mathbf{X} = (\Omega, (\mathcal{F}_t)_{t\in I}, \mathbb{P}, (X_t)_{t\in I})$ that evolve in $\mathbb{R}^d$. The main contribution of this article is to provide for every $r \geq 0$ an explicit map

$$\Phi_r : \mathcal{S} \to \mathbf{T}^{r+1}, \quad \mathbf{X} \mapsto \Phi_r(\mathbf{X})$$

from $\mathcal{S}$ into a normed and graded space[3] $\mathbf{T}^{r+1}$ such that the adapted topology of rank r on $\mathcal{S}$, τ_r, arises as the *initial topology* for Φ_r. That is τ_r is the coarsest topology τ on $\mathcal{S}$ that makes the map

$$\Phi_r : (\mathcal{S}, \tau) \to (\mathbf{T}^{r+1}, \|\cdot\|)$$

continuous. Equivalently, the adapted topology of rank r, τ_r, is characterized by the universal property that any map f from a topological space into $\mathcal{S}$ is continuous if and only if $\Phi_r \circ f$ is continuous. We highlight three consequences of this result:

Metrizing adapted topologies of any rank r. It immediately follows that

$$\mathcal{S} \times \mathcal{S} \to [0, \infty), \quad (\mathbf{X}, \mathbf{Y}) \mapsto \|\Phi_r(\mathbf{X}) - \Phi_r(\mathbf{Y})\| \tag{4.2}$$

is a semi-metric on $\mathcal{S}$ that induces τ_r. For $r = 0$ and general stochastic processes our results reduce to the previously known result [46] that the expected signature map Φ_0 can metrize weak convergence; for $r = 1$ this adds a novel entry to the list of semi-metrics that induce τ_1, see [6]; for $r \geq 2$ this (semi-)metric seems to be the first constructive and computable metrization[4] of $(\mathcal{S}, \tau_r)$. Further, our results are not restricted to processes equipped with their natural filtration.

Dynamic Programming. For $r = 0$, the map Φ_0 reduces to the expected signature map. A direct application of dynamic programming shows that for a Markov process $\mathbf{X}$, $\Phi_0(\mathbf{X})$ and consequently the semi-metric (4.2), can be efficiently computed by dynamic programming. For $r \geq 1$, the maps Φ_r are constructed by recursion and we show this can be used to bootstrap dynamic programming principles, so that for any $r \geq 0$ the map Φ_r resp. the semi-metric (4.2) can be efficiently computed for Markov processes.

[3] Φ_r maps into $\mathbf{T}^{r+1}$ and notationally it would be nicer to rename $\mathbf{T}^{r+1}$ as $\mathbf{T}^r$; however, for $r = 1$ this would clash with the conventional use that $\mathbf{T}^1$ is the classical tensor algebra $\mathbf{T}$.

[4] Analogous, to how the moment sequence of a vector-valued random variable yields a metric for measures on $\mathbb{R}^d$; see Section 4.3.1. We also draw attention to [10].

A multi-graded "feature map". The maps Φ_r embed a stochastic process into linear spaces $\mathbf{T}^{r+1}$ that arise via a classic free construction in algebra, namely *the free algebra functor*. In particular, $\mathbf{T}^{r+1}$ has a natural multi-grading and $\Phi_r(\mathbf{X})$ use this to describe the interplay of the law and the filtration of the process $\mathbf{X}$ in a hierarchical manner; analogous to how the classical moments of a vector-valued random variable is graded by the moment degree.

We believe the last point is the strongest contribution of this approach to the existing literature since the embedding

$$\mathbf{X} \to \Phi_r(\mathbf{X})$$

of an adapted process $\mathbf{X}$ into a multi-graded linear space $(\mathbf{T}^{r+1}, \|\cdot\|)$ delivers more than a semi-metric. This seems to be novel even for the well-studied case of $r = 1$. For example, for $r = 0$, Φ_0 is just the expected signature map and many recent applications in statistics, machine learning and finance rely on the co-ordinates and the grading of $\Phi_0(\mathbf{X})$. In Section 4.5 we give a simple supervised classification example that demonstrates how expected signatures as they are currently used in machine learning, i.e. Φ_0, can yield a too coarse description even for simple Markov processes and how this is resolved by Φ_r for $r \geq 1$. We also mention that adapted topologies (so far, via causal Wasserstein semi-metric) are finding applications in machine learning, see [204], and the use of Φ_r in this context seems to be interesting future research venue.

Remark 4.1.1. We focus on finite discrete time processes for two reasons:

(i) Most applications and in fact, much of the recent literature on adapted topologies, studies finite discrete time.

(ii) The resulting signature and tensor structure that capture filtrations are already novel and interesting to study in finite discrete time. Some definitions and results immediately extend to continuous time, but others lead quickly to challenging research programmes; e.g. for $r = 1$ the prediction process $t \mapsto \hat{\mathbf{X}}^1_t = \mathbb{P}(X \in \cdot|\mathcal{F}_t)$ has only càdlàg trajectories, even if the sample paths of $t \mapsto X_t$ are continuous. Càdlàg rough path theory is an area of ongoing research [42, 77] and the question of how tightness propagates through such iterated (higher rank) constructions seems hard due to a lack of Prohorov type results; see Section 4.4.3 for details.

Remark 4.1.2. Our results are not restricted to stochastic processes evolving in compact subsets of finite-dimensional state spaces discussed above. By using robust

signatures [46] adapted processes that evolve in general separable Banach space are included in our approach. In this non-compact case, it turns out that the Hoover–Keisler approach of specifying an adapted topology via AF_r and the natural generalization of Aldous's approach given by iterating prediction process yield in general different topologies which might be of independent interest.

4.1.3 Outline and Notation.

The rest of the paper is laid out as follows:

- Section 4.2 recalls Hoover–Keisler's *adapted functionals* $\mathrm{AF} = \bigcup_{r\geq 0} \mathrm{AF}_r$ and the adapted topology of rank r, τ_r. Further, it identifies Aldous prediction $\hat{X}^1$ as the rank $r = 1$ construction in the sequence of rank r prediction process that we define as
$$\hat{X}_t^{r+1} := \mathbb{P}(\hat{X}^r \in \cdot | \mathcal{F}_t), \quad \hat{X}^0 := X.$$
These prediction processes evolve in state spaces that have a rich structure; e.g.
$$\begin{aligned}
&\mathrm{Law}(\hat{\mathbf{X}}^0) \in \mathrm{Meas}(I \to V) =: \mathcal{M}_1,\\
&\mathrm{Law}(\hat{\mathbf{X}}^1) \in \mathrm{Meas}(I \to \mathrm{Meas}(I \to V)) =: \mathcal{M}_2,\\
&\mathrm{Law}(\hat{\mathbf{X}}^3) \in \mathrm{Meas}(I \to \mathrm{Meas}(I \to \mathrm{Meas}(I \to V))) =: \mathcal{M}_3.
\end{aligned}$$
We refer to the spaces $\mathcal{M}_r$ as *rank r measures*. Capturing their structure is the central theme of this article.

- Section 4.3 discusses how an element of $\mathcal{M}_r$ can be described by a multi-graded sequence of tensors. For $r = 0$, we recall that the signature S_1 injects a path into the free algebra $\mathbf{T}^1$ that consists of sequences of tensors of increasing degree; the expected signature $\bar{\mathrm{S}}_1$ injects $\mathcal{M}_1$ into $\mathbf{T}^1$. To generalize this from $r = 1$ to general $r \geq 1$ we first introduce the space of *higher rank paths* V_r: for a linear space V define
$$V_{r+1} := I \to V_r \quad V_0 := V.$$
The *rank r signature* $\mathrm{S}_r : V_r \to \mathbf{T}^r$ then injects a rank r path into the *rank r tensor algebra* $\mathbf{T}^r$ which consists of sequences of multi-graded sequences of tensors; the *rank r expected signature* $\bar{\mathrm{S}}_r : \mathcal{M}_r \to \mathbf{T}^r$ provides a multi-graded description of an element of $\mathcal{M}_r$ by injecting it into $\mathbf{T}^r$.

- Section 4.4 contains our main theoretical results. We first show that convergence in the adapted topology τ_r is equivalent to convergence in law of the rank r prediction process. This allows us to show that the rank $r+1$ expected signature $\bar{\mathrm{S}}_{r+1}$ applied to the rank r prediction process induces the rank r topology τ_r. Hence, the map

$$\Phi_r : \mathcal{S} \to \mathbf{T}^{r+1}, \quad \Phi_r(\mathbf{X}) := \bar{\mathrm{S}}_{r+1}(\mathrm{Law}(\hat{X}^r))$$

  induces τ_r as initial topology on $\mathcal{S}$.

- Section 4.5 shows that the maps $\Phi_r(\mathbf{X})$ can be efficiently computed by dynamic programming when $\mathbf{X}$ is a Markov process. We provide a Python implementation[5] of the resulting algorithms and use it for a simple numerical experiment that demonstrates the advantages of Φ_1 against the usual expected signature Φ_0.

- Appendix 4.A contains details for Example 4.1.2, Appendix 4.B contains some details on the construction of higher rank tensor algebras, and Appendix 4.B.4 contains some background on the robust signature and how it can be used to overcome problems arising from non-compactness.

Symbol	Meaning	Page
	Spaces	
V	a separable Banach space	125
U	a topological space	121
$\mathcal{S}(\mathrm{U})$	the set of adapted stochastic processes in U	121
$\underline{\Omega}$	an adapted probability space $\underline{\Omega} = (\Omega, \mathbb{P}, (\mathcal{F}_t))$	120
$\mathbf{X} = (\underline{\Omega}, X) \in \mathcal{S}(\mathrm{U})$	an adapted process on the stochastic base $\underline{\Omega}$	122
$\mathrm{Meas}(\mathrm{U})$	Borel measures on U	124
$\mathrm{Prob}(\mathrm{U})$	Borel probability measures on U	125
I	A finite totally ordered set (time)	113
$\big(I \to \mathrm{U}\big)$	the space of sequences in U indexed by I	127
	The Adapted Topology of Rank r	
AF	adapted functionals, $f(\mathbf{X})$ is a real-valued random variable	115
$\mathrm{AF}_r = \{f \in \mathrm{AF} \mid \mathrm{rank}(f) \leq r\}$	adapted functionals with rank less than r	115
τ_r	the adapted topology of rank r on $\mathcal{S}(\mathrm{U})$	123
$\hat{\tau}_r$	the extended weak topology of rank r on $\mathcal{S}(\mathrm{U})$	123

[5] Available at https://github.com/PatricBonnier/Higher-rank-signature-regression

Paths and Measures of Rank r		
U_r	the space of rank r paths with state space U	129
$\mathcal{M}_r(\mathrm{U})$	the space of rank r Borel measures on U	131
$\mathcal{P}_r(\mathrm{U})$	the space of rank r Borel probability measures on U	131
(Expected) Signature of Rank r		
$\mathbf{T}^r(V)$	The rank r tensor algebra	129
$\hat{\mathbf{X}}^r$	the rank r-prediction process $\hat{\mathbf{X}}^r_t := \mathbb{P}(\hat{\mathbf{X}}^{r-1} \in \cdot \vert \mathcal{F}_t)$ of $\mathbf{X} \in \mathcal{S}$	122
S_r	the rank r signature map $\mathrm{S}_r : V_r \to \mathbf{T}^r(V)$	129
$\bar{\mathrm{S}}_r$	the rank r expected signature map $\bar{\mathrm{S}}_r : \mathcal{M}_r(V) \to \mathbf{T}^r(V)$	132
$\bar{\mathbf{X}}^r$	the rank r conditional expected signature $\bar{\mathbf{X}}^r_t := \mathbb{E}[\bar{\mathbf{X}}^{r-1} \vert \mathcal{F}_t]$	136
$d_r(\mathbf{X}, \mathbf{Y})$	the rank r adapted signature distance between $\mathbf{X}$ and $\mathbf{Y}$	134

4.2 The Adapted Topology τ_r and the Extended Weak Topology $\hat{\tau}_r$

In this section we recall work of Hoover–Keisler [101], and define adapted functionals AF_r of rank r. We then revisit Aldous [3] notion of a prediction process, and generalize it to rank r prediction processes; that is we associate with every element $\mathbf{X} \in \mathcal{S}(\mathrm{U})$ the sequence $(\hat{\mathbf{X}}^r)_{r \geq 0}$ of rank r prediction processes. Both of the resulting objects – adapted functionals of rank r resp. prediction processes of rank r – can be used to define a topology on the space $\mathcal{S}(\mathrm{U})$ of adapted processes with state space U and intuitively capture more structural information of the filtration as r increases. We refer to these two topologies as the adapted topology τ_r of rank r and the extended weak topology of rank r.

Definition 18. Denote by $I = \{0, 1, \dots, T\}$. A filtered probability space is a triple $\underline{\Omega} = (\Omega, \mathbb{P}, (\mathcal{F})_{t \in I})$ consisting of a sample space Ω, a probability measure $\mathbb{P}$, and a filtration $(\mathcal{F}_t)_{t \in I}$. An adapted stochastic process $\mathbf{X} = (\underline{\Omega}, X)$ with state space U consists of a filtered probability space $\underline{\Omega}$ and a map $X : \Omega \times I \to \mathrm{U}$ such that X_t is $\mathcal{F}_t$-measurable for each $t \in I$. Denote with $\mathcal{S}(\mathrm{U})$ the space of adapted stochastic

processes that evolve in discrete time I in a state space U,

$$\mathcal{S}(\mathrm{U}) := \{\mathbf{X} \,|\, \mathbf{X} \text{ is an adapted stochastic process indexed by } I \text{ that evolves in U}\}.$$

We also set

$$\mathrm{Law}(\mathbf{X}) := \mathbb{P} \circ X^{-1} \text{ for } \mathbf{X} = (\Omega, \mathbb{P}, (\mathcal{F})_{t\in I}, X).$$

With the usual slight abuse of notation we use throughout the same symbol $\mathbb{E}$ for the expectation although the elements of $\mathcal{S}(\mathrm{U})$ can be supported on different adapted probability spaces.

4.2.1 Adapted Functionals

A natural way to define a topology on $\mathcal{S}(\mathrm{U})$ is by specifying some set of functionals and requiring that

$$\mathbf{X}_n \to \mathbf{X} \text{ if } \mathbb{E}[f(\mathbf{X}_n)] \to \mathbb{E}[f(\mathbf{X})] \text{ as } n \to \infty$$

for every f in this set of functionals. By choosing the set of functionals to be

$$\{f \,|\, f(\mathbf{X}) = g(X_{t_1}, \ldots, X_{t_n}),\ g \in C_b(\mathrm{U}^n, \mathbb{R}),\ (t_1, \ldots, t_n) \in I^n\}$$

one recovers classical weak convergence. In view of the above examples, it is natural to construct a wider class of functionals by using the conditional expectation in order to capture some of the information contained in the filtration.

Definition 19. We define a set of maps AF from $\mathcal{S}(\mathrm{U})$ into the set of real-valued random variables inductively:

1. if $t_1, \ldots, t_n \in I$ and $f \in C_b(\mathrm{U}^n, \mathbb{R})$, then $\mathbf{X} \mapsto f(X(t_1), \ldots, X(t_n)) \in \mathrm{AF}$,
2. if $f_1, \ldots, f_n \in \mathrm{AF}$ and $f \in C_b(\mathbb{R}^n, \mathbb{R})$, then $\mathbf{X} \mapsto f(f_1(\mathbf{X}), \ldots, f_n(\mathbf{X})) \in \mathrm{AF}$,
3. if $f \in \mathrm{AF}$ and $t \in I$ then $\mathbf{X} \mapsto \mathbb{E}[f(\mathbf{X})|\mathcal{F}_t] \in \mathrm{AF}$.

We refer to the elements of AF as adapted functionals[6].

Remark 4.2.1. For a given $\mathbf{X} = (\Omega, \mathbb{P}, (\mathcal{F})_{t\in I}, X)$ and $f \in \mathrm{AF}$, $f(\mathbf{X})$ is in $L^\infty(\Omega, \mathbb{P})$, hence the image set of $f \in \mathrm{AF}$ is $\prod_{\mathbf{X}\in\mathcal{S}(\mathrm{U})} L^\infty(\Omega^{\mathbf{X}}, \mathbb{P}^{\mathbf{X}})$ where we write $\mathbf{X} = (\Omega^{\mathbf{X}}, \mathbb{P}^{\mathbf{X}}, (\mathcal{F}^{\mathbf{X}})_{t\in I}, X)$ to emphasize the dependence of the underlying filtered probability spaces on $\mathbf{X}$.

[6] In [101] AF are called conditional processes.

Intuitively, the more times the conditional expectation is iterated the more of the evolutional constraints that are encapsulated in the filtration are exposed by the functionals in AF. Indeed, Figure 4.1 shows two processes that can not be distinguished without at least one iteration, and in Example 4.1.2, at least two iterations are required. With this in mind, we define the rank r of an adapted functional $f \in \mathrm{AF}$ as the minimal number of times the conditional expectation is iterated in the construction of f. This number r of conditional expectations gives AF a natural grading.

Definition 20. Define $\mathrm{rank} : \mathrm{AF} \to \mathbb{N} \cup \{0\}$ as

1. $\mathrm{rank}(f) = 0$ if $f(\mathbf{X}) = g(X_{t_1}, \ldots, X_{t_n})$ for $g \in C_b(\mathrm{U}^n, \mathbb{R})$
2. $\mathrm{rank}(f) = \max(\mathrm{rank}(f_1), \ldots, \mathrm{rank}(f_n))$ if $f(\mathbf{X}) = g(f_1(\mathbf{X}), \ldots, f_n(\mathbf{X}))$, $g \in C_b(\mathbb{R}^n, \mathbb{R})$, $f_1, \ldots, f_n \in \mathrm{AF}$,
3. $\mathrm{rank}(f) = \mathrm{rank}(g) + 1$ if $f(\mathbf{X}) = \mathbb{E}[g(\mathbf{X})|\mathcal{F}_t]$ for $g \in \mathrm{AF}$.

We call

$$\mathrm{AF}_r := \{f \in \mathrm{AF} | \, \mathrm{rank}(f) \leq r\}$$

the set of adapted functionals of rank less than r.

Remark 4.2.2. Following Definition 19, every $f \in \mathrm{AF}$ can be obtained by repeating steps 1, 2 and 3 finitely many times. Let $\mathfrak{r}_f$ denote such an iterative procedure which leads to the construction of $f \in \mathrm{AF}$, and let $|\mathfrak{r}_f|$ denote the total number of times step 3 (taking conditional expectation) appears in $\mathfrak{r}_f$. Note that f does not uniquely determine $\mathfrak{r}_f$; for instance, $f(\mathbf{X}) = g(X_{t_1}, \ldots, X_{t_n}) = \mathbb{E}[g(X_{t_1}, \ldots, X_{t_n})|\mathcal{F}_T]$ holds for all $\mathbf{X} \in \mathcal{S}(\mathrm{U})$ if g is a constant function. So, strictly speaking, the map $\mathrm{rank}(f)$ is given by $\mathrm{rank}(f) := \min\{|\mathfrak{r}_f| : \mathfrak{r}_f \text{ is a representation of f}\}$. However, the above (strictly speaking, not well-defined) Definition 20 is more intuitive.

4.2.2 Prediction Processes of Rank r

We now revisit Aldous' notion of prediction process. By introducing "prediction processes of prediction processes" one arrives at another natural sequence of objects (prediction processes of rank r) that capture more structure of the filtration.

Definition 21. Let $\mathbf{X} = (\underline{\Omega}, X) \in \mathcal{S}(\mathrm{U})$. The adapted stochastic processes $(\hat{\mathbf{X}}^r)_{r \geq 0}$ of $\mathbf{X}$ are defined as $\hat{\mathbf{X}}^r = (\underline{\Omega}, \hat{X}^r)$ with $\hat{X}^r$ given inductively as

$$\hat{X}^0 := X \text{ and } \hat{X}^{r+1} := (\mathbb{P}(\hat{X}^r \in \cdot | \mathcal{F}_t))_{t \in I}.$$

We call $\hat{\mathbf{X}}^r$ the rank r prediction process of $\mathbf{X}$ and we denote with U_r the state space of the process $\hat{X}^r$.

An immediate but useful identity that we use several times is that

$$\hat{X}_0^r = \mathrm{Law}(\hat{X}^{r-1}).$$

4.2.3 The Adapted and the Weak Extended Topology of Rank r

We now have two natural ways to generalize the definition of weak convergence so that it takes the filtration into account: one by replacing continuous bounded functions by adapted functions; one by replacing weak convergence of the process by weak convergence of the prediction process.

Definition 22. Let $r \geq 0$. We say that two adapted processes $\mathbf{X} \in \mathcal{S}(\mathrm{U})$ and $\mathbf{Y} \in \mathcal{S}(\mathrm{U})$ have the same adapted distribution up to rank r, in notation $\mathbf{X} \equiv_r \mathbf{Y}$, if

$$\mathbb{E}[f(\mathbf{X})] = \mathbb{E}[f(\mathbf{Y})] \quad \forall f \in \mathrm{AF}_r.$$

Moreover, we say that a sequence $(\mathbf{X}^n)_{n\geq 0} \subset \mathcal{S}(\mathrm{U})$ converges to $\mathbf{X} \in \mathcal{S}(\mathrm{U})$ in

1. the extended weak topology of rank r if

$$\lim_{n\to\infty} \mathbb{E}[f(\hat{\mathbf{X}}^{r,n})] = \mathbb{E}[f(\hat{\mathbf{X}}^r)] \quad \forall f \in C_b(\mathrm{U}_r, \mathbb{R})$$

   where U_r denotes the state space of process $\hat{\mathbf{X}}^r$.

2. the adapted topology of rank r if

$$\lim_{n\to\infty} \mathbb{E}[f(\mathbf{X}_n)] = \mathbb{E}[f(\mathbf{X})] \quad \forall f \in \mathrm{AF}_r.$$

The extended weak topology on $\mathcal{S}(\mathrm{U})$ is denoted by $\hat{\tau}_r$ and the adapted topology of rank r by τ_r.

In Section 4.4 we show that

$$(\mathcal{S}(\mathrm{U}), \tau_r) = (\mathcal{S}(\mathrm{U}), \hat{\tau}_r)$$

whenever U is compact but that for non-compact subsets U of Banach spaces, τ_r is in general coarser than $\hat{\tau}_r$; that is $\tau_r \subsetneq \hat{\tau}_r$.

4.3 (Expected) Signatures of Rank r

In the previous Section 4.2 we have introduced two topologies on $\mathcal{S}(\mathrm{U})$, τ_r and $\hat{\tau}_r$. We expect both to capture more or less the same structure (except for some subtle issues when U is non-compact). However, an attractive property of the extended weak topology $\hat{\tau}_r$ of rank r is that it is specified by classical weak convergence of a stochastic process, namely weak convergence of the rank r prediction process $\hat{\mathbf{X}}^r$. For $r = 0$ it is known that weak convergence of a stochastic processes – such as the prediction process $\hat{\mathbf{X}}^0$ – can be characterized as convergence of the expected signatures, [46]. This suggests that a similar approach can be fruitful in capturing the weak convergence of the higher rank prediction processes $\hat{\mathbf{X}}^r$.

Unfortunately, for $r \geq 1$ the rank r prediction processes evolve in very large state spaces (of laws) that have a rich and nested structure which makes the use of expected signatures less straightforward. In this section we introduce higher rank (expected) signatures that are capable of capturing the law of such processes and their nested structure. The key is to think about so-called higher rank paths that arise by currying multi-parameter paths.

4.3.1 Recall: Moment Sequences of Random Variables

Before we discuss signatures it is instructive to briefly revisit classical moment sequences and fix some notation.

Moments and Duality

Recall that for any compact set $\mathrm{U} \subseteq V = \mathbb{R}^d$, the moment map

$$\mathrm{Meas}(\mathrm{U}) \hookrightarrow \prod_{m \geq 0} V^{\otimes m}, \quad \mu \mapsto \left(\int x^{\otimes m} \mu(dx) \right)_{m \geq 0} \tag{4.3}$$

is an injection from the space Meas(U) of signed Borel measures on U to $\prod_{m \geq 0} V^{\otimes m}$. A short way to prove the injection (4.3) is to recall that the dual of Meas(U) is the space of $C_b(\mathrm{U}, \mathbb{R})$. Under this duality, injectivity of the map (4.3) amounts to the density of monomials in $C_b(\mathrm{U}, \mathbb{R})$ and the latter follows immediately by the Stone–Weierstrass Theorem. Although this is not how the proof that moments can characterize laws is usually presented, this approach is very powerful when one tries to develop a similar argument on non-compact spaces, see [46]. This duality is also the main reason to work with the linear space Meas(U) although we are ultimately interested in the convex set Prob(U) of probability measures.

In particular, when restricted to the set of probability measures Prob(U) $\subset$ Meas(U), the injection (4.4) shows that the law of a U-valued random variable X, $\mu(\cdot) = \mathbb{P}(X \in \cdot)$, is characterized as an element of $\prod_{m\geq 0} V^{\otimes m}$,

$$\left(\mathbb{E}\left[\frac{X^{\otimes m}}{m!}\right]\right)_{m\geq 0} \in \prod_{m\geq 0} V^{\otimes m}.$$

Note that above we have included the factorial decay $m!$. This is convenient since these terms arise in the Taylor expansion of the exponential function. In the case of compactly supported random variables this does not make a difference but much more care need to be taken in the non-compact case and we return to this discussion in Section 4.3.5.

Tensors, Moments, Exponentials

The tensor exponential provides a concise, coordinate-free way of expressing moment relations.

Definition 23. Let V be a Banach space. Define the exponential map

$$\exp : V \mapsto \prod_{m\geq 0} V^{\otimes m}, \quad v \mapsto \left(\frac{v^{\otimes m}}{m!}\right)_{m\geq 0}.$$

To get used to this notation, it is instructive to apply the above exponential map with $V = \mathbb{R}^d$: spelled out in coordinates, the exponential map reduces to the usual moment map,

$$x = (x^i)_{i=1,\dots,d} \in \mathbb{R}^d \mapsto \left(\frac{x^{\otimes m}}{m!}\right) \simeq \left(\frac{x^{i_1}\cdots x^{i_m}}{m!}\right)_{1\leq i_1,\dots,i_m\leq d}.$$

Applied to a random variable X taking values in a compact subset U $\subset \mathbb{R}^d$, all the moments of X,

$$\mathbb{E}[\exp X] = (1, \mathbb{E}[X], \frac{1}{2!}\mathbb{E}[X^{\otimes 2}], \dots)_{m\geq 0} \in \prod_{m\geq 0} V^{\otimes m} \tag{4.4}$$

are given as the expected value of the $\prod_{m\geq 0} V^{\otimes m}$-valued random variable $\exp(X)$. Weak convergence is then characterized as convergence of the expected value of the tensor exponential.

Proposition 4.3.1. *Let (X_n) be a sequence of random variables that take values in a compact subset* U $\subset \mathbb{R}^d$. *Then X_n converges weakly to a random variable X if and only if*

$$\mathbb{E}[\exp X_n] \to \mathbb{E}[\exp X] \text{ as } n \to \infty$$

where convergence on $\prod_{m\geq 0}(\mathbb{R}^d)^{\otimes m}$ is defined as convergence on each degree $(\mathbb{R}^d)^{\otimes m}$.

Proof. The assumption of compact support implies tightness, hence the statement follows by Prohorov's theorem if one shows that if (X_n) converges weakly along a subsequence to Y, then Y equals X in law. But if $X_{n_k} \to Y$ weakly as $k \to \infty$, then by assumption $\lim_k \mathbb{E}[p(X_{n_k})] = \mathbb{E}[p(Y)]$ for any polynomial p. The assumption also implies that $\lim_n \mathbb{E}[p(X_n)] = \mathbb{E}[p(X)]$, hence

$$\mathbb{E}[p(X)] = \mathbb{E}[p(Y)]$$

for any polynomial p. Since polynomials are dense in $C(\mathrm{U}, \mathbb{R})$, this implies that $\mathrm{Law}(X) = \mathrm{Law}(Y)$. □

To put the above into the context of the rest of this paper, note that these results can be reformulated as saying that the topology of weak convergence on the space of random variables that take values in a compact state space $\mathrm{U} \subset \mathbb{R}^d$ is the initial topology of the map

$$(\Omega, \mathcal{F}, X) \mapsto \varphi\left((\Omega, \mathcal{F}, X)\right) := \left(\mathbb{E}\left[\frac{X^{\otimes m}}{m!}\right]\right)_{m \geq 0} \in \prod_{m \geq 0} V^{\otimes m}, \tag{4.5}$$

resp. the weak topology is induced by the (semi-)metric

$$(\Omega_1, \mathcal{F}, X) \times (\Omega_2, \mathcal{G}, Y) \mapsto \|\varphi((\Omega_1, \mathcal{F}, X)) - \varphi((\Omega_2, \mathcal{G}, Y))\|.$$

The space $\prod_{m \geq 0} V^{\otimes m}$ appeared above simply since it is the natural state space for the moment sequence and so far we have only cared about its linear structure since we took expectations. However, a more abstract and algebraic way to think about $\prod_{m \geq 0} V^{\otimes m}$ is that it is the free algebra of the vector space: put briefly, given a linear space V we want to embed V into a bigger space that is not only a vector space but also allows to multiply elements, that is an algebra; further we want to do this into "most general way". A classical result from algebra shows that for any Banach V space such an algebra exists and it is given as $\prod_{m \geq 0} V^{\otimes m}$. The formal statement is that the map $V \mapsto \prod_{m \geq 0} V^{\otimes m}$ from the category of Banach spaces to the category of algebras is functorial and free; see [117, Chapter 16]. So far we have not used this algebra structure, but we are about derive the analogous statement to (4.5) for stochastic processes. To do this, this additional product structure on $\prod_{m \geq 0} V^{\otimes m}$ becomes crucial. The first step is to find a suitable replacement for the tensor exponential exp to lift a path into $\prod_{m \geq 0} V^{\otimes m}$ – this leads to the notion of the signature.

4.3.2 Signatures as Non-Commutative Exponentials

To apply a similar reasoning to paths rather than vectors, one needs to take the sequential order into account as time progresses. To do so we use that the linear space of tensors, $\prod_{m\geq 0} V^{\otimes m}$, carries a natural non-commutative product.

Definition 24. For $s = (s_m)_{m\geq 0}, t = (t_m)_{m\geq 0} \in \prod_{m\geq 0} V^{\otimes m}$ define

$$s \cdot t := \left(\sum_{i=0}^{m} s_i t_{m-i} \right)_{m\geq 0} \in \prod_{m\geq 0} V^{\otimes m}. \tag{4.6}$$

We refer to $s \cdot t$ as as the so-called tensor convolution product of s and t.

To account of the sequential order in a path $x(0), x(1), \ldots, x(T)$, we now simply lift the increment of a path $x(t+1)-x(t)$ at time t into $\prod_{m\geq 0} V^{\otimes m}$ via $\exp(x(t+1)-x(t))$, and then use the tensor convolution product (4.6) to "stitch these lifted increments together". For our purposes it turns out to be useful to first augment a path with an additional time-coordinate, that is instead of increments

$$x(t+1) - x(t) \in V$$

we consider increments

$$\Delta_t x := (t+1, x(t+1)) - (t, x(t)) = (1, x(t+1) - x(t)) \in \mathbb{R} \oplus V$$

and use the tensor exponential to embed these increments into $\prod_{m\geq 0}(\mathbb{R} \oplus V)^{\otimes m}$ (the free algebra over the linear space $\mathbb{R} \oplus V$). A final but important observation is that it is better to work with a slightly smaller space $\mathbf{T}^1(V) \subset \prod_{m\geq 0}(\mathbb{R} \oplus V)^{\otimes m}$. The main reason is that on this smaller space one can canonically lift a norm on V to a norm on $\mathbf{T}^1(V)$; see Appendix 4.B and 4.C for details on $\mathbf{T}^1(V)$.

Putting everything together results in the definition of the (discrete time) signature,

Definition 25. Let V be a Banach space and $I = \{0, 1, \ldots, T\}$. The (rank 1) signature map is defined as

$$\mathrm{S} : (I \to V\Big) \to \mathbf{T}^1(V), \quad x \mapsto \prod_{t\in I} \exp \Delta_t x.$$

where $\Delta_0 x := (1, x(0))$ and $\Delta_t x := (1, x(t) - x(t-1)) \in \mathbb{R} \oplus V$. The (rank 1) expected signature map is defined as

$$\bar{\mathrm{S}} : \mathrm{Meas}(I \to V) \to \mathbf{T}^1(V), \quad \mu \mapsto \int_{x\in V^I} \mathrm{S}(x)\mu(dx).$$

A guiding principle is that the signature of a path is the natural generalization of the monomials of a vector; resp. the expected signatures of a stochastic process is the natural generalization of the moment sequence of a vector-valued random variable. Indeed, by taking $|I| = 2$, the above equation recovers the classical tensor exponential exponential (with one additional coordinate for time),

$$\mathrm{S}(x_1, x_0) = \exp(\Delta x) = \exp((x_1 - x_0, 1)) = 1 + \Delta x + \frac{1}{2}(\Delta x)^{\otimes 2} + \frac{1}{6}(\Delta x)^{\otimes 3} + \cdots .$$

From this point of view, the following Theorem is then not surprising.

Theorem 4.3.1. *Let V be a Banach space and $\mathrm{U} \subset V$ compact.*

1. *The map $\mathrm{S} : \big(I \to V\big) \to \mathbf{T}^1(V)$ is injective.*
2. *The family of linear signature functionals*

$$\{x \mapsto \langle l, \mathrm{S}(x)\rangle : \ l \in \bigoplus_{m\geq 0} (V^{\otimes m})^\star\}$$

   is dense in $\mathrm{C}(I \to \mathrm{U}, \mathbb{R})$ with the uniform norm.
3. *The map $\bar{\mathrm{S}} : \mathrm{Meas}(I \to \mathrm{U}) \to \mathbf{T}^1(V)$ is injective.*

Proof. Item 1 is a special case of [41, Theorem 1]. Item 2 is due to Fliess [72, Corollary 4.9]. Item 3 follows since if $\mu, \nu \in \mathrm{Meas}(I \to \mathrm{U})$ are such that $\bar{\mathrm{S}}(\mu) = \bar{\mathrm{S}}(\nu)$, then $\langle \bar{\mathrm{S}}(\mu), \ell\rangle = \langle \bar{\mathrm{S}}(\nu), \ell\rangle$ for any $\ell \in \bigoplus_{m\geq 0}(V^{\otimes m})^\star$ so by Item 2 it holds that $\mu(f) = \nu(f)$ for any $f \in \mathrm{C}(I \to \mathrm{U}, \mathbb{R})$, hence $\mu = \nu$. □

Remark 4.3.1. Everything in this section is classical: our discrete signature coincides with Chen's [39] iterated integral signature, that is $\mathrm{S}(x)_m = \int dx^L_{t_1} \otimes \cdots \otimes dx^L_{t_m}$ where $x : [0, T] \to V$ denotes the path given by linear interpolation of $\{(t, x(t)) : t \in I\}$. Usually, signatures are defined without the time–coordinate and only capture the path up to re-parametrization, but the adapted topologies depend on the parametrization so it is natural to include the time–coordinate. Nevertheless, the results in the following sections can be easily adapted without the additional time-coordinate and it might be interesting to study the resulting adapted topology for equivalence classes of unparametrized paths; see [46] for a discussion for the case of weak convergence, $r = 0$. See also Appendix 4.C for more on signatures.

4.3.3 Paths and Signatures of Higher Rank

Section 4.3.2 recalled that the [expected] signature can characterize [measures on] paths. Our goal is to characterize the predictions processes introduced in Section 4.2. Simply applying the expected signature to a prediction process would ignore the nested structure of the state spaces of these processes, see Definition 21, and we heavily use this structure in the proof of our main result, Section 4.4. To address this we first introduce higher rank paths which formalize paths evolving in spaces of paths and then use this to introduce higher rank [expected] signatures.

Definition 26. Let $(I_r)_{r\geq 1}$ be a sequence of finite ordered sets and U a topological space. Let $\mathrm{U}_0 := \mathrm{U}$ and define $(\mathrm{U}_r)_{r\geq 0}$ inductively,

$$\mathrm{U}_r := \Big(I_r \to \mathrm{U}_{r-1}\Big)$$

We refer to an element of U_r as a path of rank r in the state space U

Explicitly, these spaces can be unravelled as

$$\mathrm{U}_r = I_r \to \mathrm{U}_{r-1} = (I_r \to (I_{r-1} \to \cdots (I_2 \to \underbrace{(I_1 \to \mathrm{U})}_{\mathrm{U}_1})\cdots)).$$
$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{\mathrm{U}_2}$$
$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{\mathrm{U}_{r-1}}$$

A rank 1 path coincides with the usual definition of a path from I_1 into U. Evaluating a rank r path at time $t_r \in I_r$ yields a rank $r-1$ path in the same state space, that is for $x \in \mathrm{U}_r$, $x(t_r) \in \mathrm{U}_{r-1}$ for every $t_r \in I_r$.

Recall that the Signature from Definition 25 injects any path evolving in a Banach space V into the Banach space $\mathbf{T}^1(V)$. By iterating compositions of this map and defining $\mathbf{T}^{r+1}(V)$ inductively as the completion of $\bigoplus_{n\geq 0} \mathbf{T}^r(V)^{\otimes n}$ with respect to a suitable norm (see Appendix 4.B for details) we get the following.

Definition 27. Let V be a Banach space. Define the family of maps $(\mathrm{S}_r)_{r\geq 1}$,

$$\mathrm{S}_r : V_r \to \mathbf{T}^r(V)$$

inductively by setting $\mathrm{S}_1 := \mathrm{S}$ and for $r \geq 2$,

$$\mathrm{S}_r(x) := \mathrm{S}(x^\star \mathrm{S}_{r-1}),$$

where $x^\star \mathrm{S}_{r-1}$ denotes the pullback[7] of S_{r-1} by x. We call S_r the signature map of rank r.

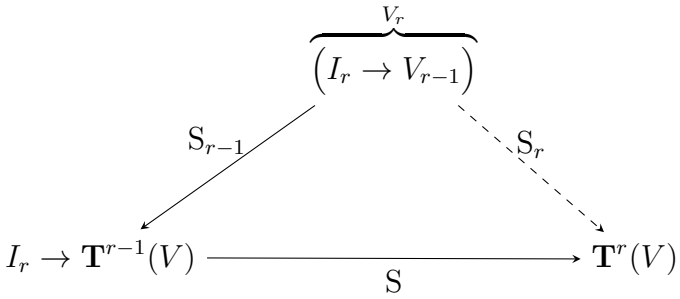


Figure 4.2: The inductive definition of S_r. By extending the map S_{r-1} to a map $V_r \to (I_r \to \mathbf{T}^{r-1}(V))$, the signature S can be applied to it to form $\mathrm{S}_r : V_r \to \mathbf{T}^r(V)$.

Example 4.3.1. Schematically, we can think of rank r signatures and rank r paths as

$$V_r = (I_r \to (I_{r-1} \to \cdots (I_2 \to \underbrace{(I_1 \to V)}_{V_1 \hookrightarrow \mathbf{T}^1(V)}) \cdots)).$$
$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{V_2 \hookrightarrow \mathbf{T}^2(V)}$$
$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{V_{r-1} \hookrightarrow \mathbf{T}^{r-1}(V)}$$

The above construction starts with applying the usual signature to the innermost bracket to turn the map $V_1 \to I_1$ into an element of $\mathbf{T}^1(V)$ (the top curly bracket). The next step turns the map $I_2 \to \mathbf{T}^1(V)$ into an element of $\mathbf{T}^2(V)$, etc. It is instructive to go through a couple of case for r.

1. For $r = 1$, we are given a (rank 1) path $x : I_1 \to V$, and $\mathrm{S}_1(x)$ is by definition the signature of x, $\mathrm{S}_1(x) \in \mathbf{T}^1(V)$. That is, S_1 maps rank 1 paths in the state space V to elements of $\mathbf{T}^1(V)$.

2. For $r = 2$, we are given a rank 2 path x in the state space V, $x : I_2 \to (I_1 \to V)$. The evaluation of x at any $t_2 \in I_2$ yields a rank 1 path in the state space V

$$x(t_2) : I_1 \to V, \quad t_1 \mapsto x(t_2)(t_1).$$

   Since S_1 maps a rank 1 path in the state space V to an element of $\mathbf{T}^1(V)$, the pullback of S_1 by $x^\star$ equals

$$x^\star \mathrm{S}_1 : I_2 \to \mathbf{T}^1(V), \quad t_2 \mapsto \mathrm{S}_1(x(t_2)).$$

[7] that is $(x^\star \mathrm{S}_{r-1})(t) := \mathrm{S}_{r-1}(x(t))$ using that $x \in V_r$ and $x(t) \in V_{r-1}$.

By definition, $\mathrm{S}_2(x)$ is the signature of this rank 1 path, $x^\star \mathrm{S}_1$, that evolves in the state space $\mathbf{T}^1(V)$,

$$\mathrm{S}_2(x) = \mathrm{S}(x^\star \mathrm{S}_1),$$

and therefore $\mathrm{S}_2(x) \in \mathbf{T}^2(V) \subseteq \prod_{m\geq 0}(\mathbb{R} \oplus \mathbf{T}^1(V))^{\otimes m}$. That is, S_2 maps rank 2 paths in the state space V to elements of $\mathbf{T}^1(V)$.

Proposition 4.3.2. *The map* $\mathrm{S}_r : V_r \to \mathbf{T}^r(V)$ *is injective.*

Proof. Follows by iterating Theorem 4.3.1. □

4.3.4 Measures and Expected Signatures of Higher Rank

Our goal is to inject the laws of predictions processes into a normed space. Recall that the laws of prediction processes have a rich nested structure, for example

$$\begin{aligned}
&\mathrm{Law}(\hat{\mathbf{X}}^0) \in \mathrm{Meas}(I \to V) =: \mathcal{M}_1,\\
&\mathrm{Law}(\hat{\mathbf{X}}^1) \in \mathrm{Meas}(I \to \mathrm{Meas}(I \to V)) =: \mathcal{M}_2,\\
&\mathrm{Law}(\hat{\mathbf{X}}^2) \in \mathrm{Meas}(I \to \mathrm{Meas}(I \to \mathrm{Meas}(I \to V))) =: \mathcal{M}_3.
\end{aligned}$$

Capturing this nested structure is essential when discussing the adapted topologies.

Definition 28. Let $I_1, \ldots, I_r$ be finite ordered sets and U a topological space. For $r = 1$ define $\mathcal{M}_0 := \mathrm{Prob}_0 := \mathrm{U}$ and for $r \geq 1$

$$\begin{aligned}
\mathcal{M}_r(\mathrm{U}) &:= \mathrm{Meas}(I_r \to \mathcal{M}_{r-1}(\mathrm{U})),\\
\mathcal{P}_r(\mathrm{U}) &:= \mathrm{Prob}(I_r \to \mathcal{P}_{r-1}(\mathrm{U}))
\end{aligned}$$

We endow $\mathcal{M}_r(\mathrm{U})$ and $\mathcal{P}_r(\mathrm{U})$ with the natural weak topology. We refer to an element of $\mathcal{M}_r(\mathrm{U})$ as a rank r measure on U and an element of $\mathcal{P}_r(\mathrm{U})$ as a rank r probability measure on U.

Clearly, $\mathcal{P}_r(U) \subset \mathcal{M}_r(U)$ and these spaces can be written explicitly as

$$\begin{aligned}
\mathcal{M}_r(\mathrm{U}) &= \mathrm{Meas}(I_r \to \mathrm{Meas}(I_{r-1} \to \cdots \mathrm{Meas}(I_2 \to \mathrm{Meas}(I_1 \to \mathrm{U}))\cdots)), \quad (4.7)\\
\mathcal{P}_r(\mathrm{U}) &= \mathrm{Prob}(I_r \to \mathrm{Prob}(I_{r-1} \to \cdots \mathrm{Prob}(I_2 \to \mathrm{Prob}(I_1 \to V))\cdots)).
\end{aligned}$$

As mentioned in Section 4.3.1, although we are interested in the convex set of probability measures $\mathcal{P}_r$, working with the larger linear space of Borel measures $\mathcal{M}_r$ allows us to use duality arguments. We emphasize that $\mathcal{M}_r(\mathrm{U})$ is significantly

bigger than $\mathrm{Meas}(\mathrm{U}_r)$: the latter embeds into the former by taking the $r-1$ innermost measures in the parenthesis in (4.7) to be Dirac measures.

Analogous to how we iterated signature maps and tensor algebras in the previous section, we now construct expected signatures to provide an injection $\mathcal{M}_r(V) \hookrightarrow \mathbf{T}^r(V)$.

Definition 29. Let V be a Banach space. Define the family of maps $(\bar{\mathrm{S}}_r)_{r\geq 1}$ inductively by setting $\bar{\mathrm{S}}_0 := \mathrm{id}_V$ and (whenever the integral is well–defined)

$$\bar{\mathrm{S}}_r : \mathcal{M}_r(V) \to \mathbf{T}^r(V), \quad \mu \mapsto \int \mathrm{S}(x^\star \bar{\mathrm{S}}_{r-1})\mu(\mathrm{d}x),$$

where $x^\star \bar{\mathrm{S}}_{r-1}$ denotes the pullback of $\bar{\mathrm{S}}_{r-1}$ by x. We call $\bar{\mathrm{S}}_r$ the expected signature map of rank r.

The following Proposition 4.3.3 generalizes that expected signature characterizes laws of processes (Theorem 4.3.1 item 3). We postpone its proof to Theorem 4.4.3.

Proposition 4.3.3. *Let V be a separable Banach space and $\mathcal{K} \subset V$ compact. Then*

$$\bar{\mathrm{S}}_r : \mathcal{P}_r(\mathcal{K}) \to \mathbf{T}^r(V)$$

is injective.

Example 4.3.2. It is instructive to run through the first few iterations of r for $\bar{\mathrm{S}}_r$. Since one can always assume that the process $X = (X_t)_{t\in I}$ is the canonical coordinate process defined on the probability space $\Big((I \to \mathcal{M}_{r-1}(V)), \mu\Big)$, we may also write $\bar{\mathrm{S}}_r(\mu) = \mathbb{E}_\mu[S \circ \bar{\mathrm{S}}_{r-1}(X)]$.

- If $r = 1$, then for any probability measure $\mu \in \mathcal{P}_1(V) \subset \mathcal{M}_1(V) = \mathcal{M}(I \to V)$, the mapping $\bar{\mathrm{S}}_1(\mu) = \mathbb{E}_{X\sim\mu}[S(X)]$ is the expected signature of the discrete–time stochastic process $X = (X_t)_{t\in I}$ with law μ.

- If $r = 2$, then for any probability measure $\mu \in \mathcal{M}_2(V) = \mathcal{M}\Big(I \to \mathcal{M}_1(V)\Big)$, fix some stochastic process $X = (X_t)_{t\in I}$ with values in $\mathcal{M}_1(V)$ and law μ. For any $t \in I$, $X^\star \bar{\mathrm{S}}_1(t) = \bar{\mathrm{S}}_1(X(t))$ is the expected signature of $X(t)$; and hence $X^\star \bar{\mathrm{S}}_1$ can be thought of as a stochastic process taking values in the vector space $\mathbf{T}^1(V)$ and we may compute its expected signature.

  For a particular example of this, if $Z = (Z_t)_{t\in I}$ is a discrete–time process taking values in V defined on some stochastic basis $(\Omega, (\mathcal{F}_t), \mathbb{P})$, then $X_t := \mathbb{P}[Z \in \cdot | \mathcal{F}_t]$

is a regular conditional distribution of Z given $\mathcal{F}_t$. Let $\mu = \mathcal{L}(X)$ be the law of the measure-valued process X, then

$$\bar{\mathrm{S}}_2(\mu) = \mathbb{E}[S(t \mapsto \mathbb{E}[S(s \mapsto Z_s)|\mathcal{F}_t])].$$

We will give a complete description of $\bar{\mathrm{S}}_r$ for this special case in Section 4.4.

4.3.5 Non-Compactness and Robust (Expected) Signatures

Even for random variables in $\mathbb{R}^d$, elementary examples show that the sequence of moments does not characterize the law when their support is non-compact; in particular, Proposition 4.3.1 is not true without compact support. The same applies to stochastic processes and their (higher rank) expected signatures.

The "robust (Signature) Moments" construction from [46] yields an extension of the injectivity of the higher rank expected signature from the previous sections to paths in general (non-compact) Banach spaces. We emphasize that the results in Section 4.4 are already interesting for the case of compact state spaces that we have discussed in the previous sections and we invite readers less familiar with signatures to skip this section.

Proposition 4.3.4. *Let V be separable Banach space. For every $r \geq 0$ there exist maps*

$$\mathrm{S}_r^n : V_r \to \mathbf{T}^r(V)$$
$$\bar{\mathrm{S}}_r^n : \mathcal{M}_r(V) \to \mathbf{T}^r(V)$$

that are both bounded, continuous and injective. Further, the space $\mathbf{T}^r(V)$ is also a separable Banach space. We refer to S_r^n[8] as the robust signature map of rank r and to $\bar{\mathrm{S}}_r^n$ as the robust expected signature map of rank r.

Proof. The fact that every space $\mathbf{T}^r(V)$ for $r \geq 1$ is a separable Banach space follows immediately from Definition 35. Then we can show that on each $\mathbf{T}^r(V)$ there exists a so–called tensor normalization map Λ with codomain the unit ball of $\mathbf{T}^r(V)$ such that the composition $\Lambda\,\mathrm{S} := \Lambda \circ \mathrm{S}$ preserves the algebraic properties of the signature map. We refer to 4.B.4 for details on the construction of the normalization Λ. Let μ be an element of $\mathrm{Prob}_r(V) = \mathrm{Prob}(I \to \mathrm{Prob}_{r-1}(V))$. Let Z^{r-1} be a

[8]Here the superscript "n" of S_r^n means "normalized" and does not stand for the "n-th level of signature" which was often used in some literature.

stochastic process with values in $\mathrm{Prob}_{r-1}(V)$ and law μ, then we define for every $r \geq 1$, $\mathrm{S}_r^n(Z^{r-1}) := \Lambda\,\mathrm{S} \circ (t \mapsto \bar{\mathrm{S}}_{r-1}^n(Z_t^{r-1}))$ and

$$\bar{\mathrm{S}}_r^n(\mu) = \mathbb{E}_{Z^{r-1}\sim\mu}[\mathrm{S}_r^n(Z^{r-1})]$$

with the convention $\bar{\mathrm{S}}_0^n(Z^0) = Z^0$. If $r = 1$, then this follows from Proposition 4.B.3 in Appendix 4.B.4. By the induction hypothesis, $\bar{\mathrm{S}}_{r-1}^n$ is continuous and injective, hence the assertion about injectivity follows immediately from Proposition 4.C.1. Finally, the continuity of $\bar{\mathrm{S}}_r^n$ follows from the Skorokhod representation theorem since if V is separable, then $I \to V$ is separable, and hence $\mathrm{Prob}_{r-1}(V)$, is separable with respect to the weak topology. □

For ease of notation we are going to redefine the symbols S_r and $\bar{\mathrm{S}}_r$ for the remainder of the article.

Definition 30. Let V be a separable Banach space, $\mathrm{U} \subset V$, and $r \geq 0$. For the rest of this article we denote

$$\begin{aligned} \mathrm{S}_r : \mathrm{U}_r \to \mathbf{T}^r(V), &\quad x \mapsto \mathrm{S}_r^n(x), \\ \bar{\mathrm{S}}_r : \mathcal{M}_r(\mathrm{U}) \to \mathbf{T}^r(V), &\quad \mu \mapsto \bar{\mathrm{S}}_r^n(\mu). \end{aligned}$$

4.4 The Adapted Topology and Higher Rank Signatures

Our leitmotif is that expected signatures can be regarded as a generalization of the classical moment map. Indeed, for $r = 0$ we have by definition that $\mathbf{X} = \hat{\mathbf{X}}^0$ and the initial topology of the map

$$\mathcal{S} \to \mathbf{T}^1(V), \quad \mathbf{X} \to \bar{\mathrm{S}}_1(\mathrm{Law}(\hat{\mathbf{X}}^0))$$

is the topology of weak convergence $(\mathcal{S}, \tau_0)$, see [46]. This suggests that, at least locally, the initial topology of the map

$$\mathcal{S} \to \mathbf{T}^r(V), \quad \mathbf{X} \to \bar{\mathrm{S}}_{r+1}(\mathrm{Law}(\hat{\mathbf{X}}^r))$$

is the rank r adapted topology $(\mathcal{S}_r, \tau_r)$. In this Section we show that this is indeed true in great generality.

Definition 31. Let V be a separable Banach space and $\mathrm{U} \subset V$. For $r \geq 0$ define

$$\Phi_r : \mathcal{S}(\mathrm{U}) \to \mathbf{T}^{r+1}(V), \quad \mathbf{X} \mapsto \bar{\mathrm{S}}_{r+1}(\mathrm{Law}(\hat{\mathbf{X}}^r))$$

and

$$d_r : \mathcal{S}(\mathrm{U}) \times \mathcal{S}(\mathrm{U}) \to [0, \infty), \quad (\mathbf{X}, \mathbf{Y}) \mapsto \|\Phi_r(\mathbf{X}) - \Phi_r(\mathbf{Y})\|_{r+1}.$$

Our main result is

Theorem 4.4.1. *Let V be a separable Banach space and $\mathrm{U} \subset V$ compact. The following topologies on $\mathcal{S}(\mathrm{U})$ are equal*

1. *the adapted topology of rank r, τ_r,*
2. *the extended weak topology of rank r, $\hat{\tau}_r$,*
3. *the initial topology of the map $\Phi_r : \mathcal{S}(\mathrm{U}) \to \mathbf{T}^r(V)$,*
4. *the topology induced by convergence in the semi-metric d_r on $\mathcal{S}(\mathrm{U})$.*

Moreover, the same statement holds locally if U *is not compact; see Theorem 4.4.3.*

Restricted to $r = 1$ and processes with their natural filtration, the semi-metric d_1 adds another entry to the list of distances that induce the adapted topology τ_1. However, even for this $r = 1$ case, the characterization of the adapted topology as the initial topology of a map into a normed, graded space rather than the topology induced by a (semi-)metric, is to the best of our knowledge new.

Corollary 4.4.1 ([6])**.** *Let* U *be as in Theorem 4.4.1 and denote by $\mathcal{S}_{Natural}(\mathrm{U})$ the subset of $\mathcal{S}(\mathrm{U})$ of processes equipped with their natural filtration. Then the following topologies on $\mathcal{S}_{Natural}(\mathrm{U})$ are equal*

- *the topology induced by d_1,*
- *the topology induced by adapted Wasserstein distance,*
- *the topology induced by symmetrized-causal Wasserstein distance,*
- *Hellwig's information topology,*
- *Aldous' extended weak topology,*
- *the optimal stopping topology.*

The remainder of this Section is devoted to the proof of Theorem 4.4.1.

4.4.1 Higher Rank Conditional Signature Process.

The domain of $\bar{\mathrm{S}}_r$ is all of $\mathcal{M}_r(V)$. When restricted to the laws of prediction processes, this additional structure yields an useful interpretation in terms of conditional expectations; e.g. for $r = 1$ and $t \in I$,

$$\bar{\mathrm{S}}_1(\hat{\mathbf{X}}^1_t) = \int \mathrm{S}(x)\mathbb{P}[X \in dx|\mathcal{F}_t] = \mathbb{E}[\mathrm{S}(X)|\mathcal{F}_t]. \tag{4.8}$$

This motivates the following definition

Definition 32. Let $\mathbf{X} = (\Omega, (\mathcal{F}_t), \mathbb{P}, X) \in \mathcal{S}(V)$. We define a family of adapted processes $(\bar{\mathbf{X}}^r)_{r\geq 0}$ by $\bar{\mathbf{X}}^r = (\Omega, \mathcal{F}, \mathbb{P}, \bar{X}^r)$ with $\bar{X}^r$ given inductively as

$$\bar{X}^r_t := \mathbb{E}[\mathrm{S}(\bar{\mathbf{X}}^{r-1})|\mathcal{F}_t]$$

and $\bar{\mathbf{X}}^0_t = X_t$. We call $\bar{\mathbf{X}}^r$ the rank r conditional signature process of $\mathbf{X}$.

Proposition 4.4.2. *For every $r \geq 1$ and $\mathbf{X} \in \mathcal{S}(V)$ it holds that*

$$\bar{\mathrm{S}}_r(\hat{\mathbf{X}}^r_t) = \bar{\mathbf{X}}^r_t \quad \forall t \in I. \tag{4.9}$$

In particular,

$$\Phi_r(\mathbf{X}) \equiv \bar{\mathrm{S}}_{r+1}(\mathrm{Law}(\hat{\mathbf{X}}^r)) = \mathbb{E}\bar{\mathbf{X}}^{r+1}_0.$$

Proof. The second claim follows immediately from (4.9) since

$$\mathbb{E}\bar{\mathbf{X}}^r_t = \mathbb{E}\bar{\mathrm{S}}_r(\hat{\mathbf{X}}^r_t) = \mathbb{E}\int \bar{\mathrm{S}}_r(x)\mathbb{P}[\hat{\mathbf{X}}^{r-1} \in dx|\mathcal{F}_t] = \int \bar{\mathrm{S}}_r(x)\mathbb{P}[\hat{\mathbf{X}}^{r-1} \in dx] = \bar{\mathrm{S}}_r(\mathrm{Law}(\hat{\mathbf{X}}^{r-1})).$$

For the proof of (4.9) we proceed by induction over $r \geq 1$. The starting case, $r = 1$, is given in (4.8). For the induction step, assume that (4.9) holds true for some $r \geq 1$. We denote by μ_r the measure

$$\mu_r = \mathbb{P}(\hat{X}^r \in \cdot|\mathcal{F}_t).$$

By definition of $\bar{\mathrm{S}}_{r+1}$ we see that

$$\bar{\mathrm{S}}_{r+1}(\hat{\mathbf{X}}^{r+1}_t) = \int \mathrm{S}(x^\star\bar{\mathrm{S}}_r)\mu_r(dx) = \mathbb{E}[\mathrm{S}(s \mapsto \bar{\mathrm{S}}_r(\hat{X}^r_s))|\mathcal{F}_t] = \mathbb{E}[\mathrm{S}(s \mapsto \bar{\mathbf{X}}^r_s)|\mathcal{F}_t]$$

where we used the induction hypothesis, $\bar{\mathrm{S}}_r(\hat{X}^r_s) = \mathbf{X}^r_s$ in the last step. □

This interpretation of $\Phi_r(\mathbf{X})$ in terms of the rank r conditional signature process $\bar{\mathbf{X}}^{r+1}$ turns out to be very useful in the next section, in particular for the proof of Theorem 4.4.2.

4.4.2 Embedding and Metrizing Adapted Topologies

Theorem 4.4.2. *Let V be a a separable Banach space and $\mathbf{X}, \mathbf{Y} \in \mathcal{S}(V)$. For every $r \geq 0$ the following are equivalent*

1. $\mathbb{E}[f(\mathbf{X})] = \mathbb{E}[f(\mathbf{Y})] \quad \forall f \in \mathrm{AF}_r$,
2. $\mathrm{Law}(\hat{\mathbf{X}}^r) = \mathrm{Law}(\hat{\mathbf{Y}}^r)$,
3. $\mathrm{Law}(\hat{\mathbf{X}}^0, \ldots, \hat{\mathbf{X}}^r) = \mathrm{Law}(\hat{\mathbf{Y}}^0, \ldots, \hat{\mathbf{Y}}^r)$.
4. $\Phi_r(\mathbf{X}) = \Phi_r(\mathbf{Y})$.

We prepare the proof of Theorem 4.4.2 with a Lemma.

Lemma 4.4.3. *For every $r \geq 0$ and every Borel set $B \subseteq (I \to \mathcal{M}_r)$, there exists a sequence of uniformly bounded adapted functionals $f_k \in \mathrm{AF}_r$ such that $1_B \circ \hat{\mathbf{X}}^r = \lim_{k\to\infty} f_k(\mathbf{X})$ in probability.*

Proof of Lemma 4.4.3. If $r = 0$, then since V is a Polish space and $\hat{\mathbf{X}}^0 = X$, the claim holds due to Urysohn's lemma and Dynkin's lemma.

Now consider the case $r \geq 1$. Since AF_r is an algebra, by Dynkin's lemma it suffices to consider the case $B = B_0 \times \ldots \times B_T$, where each B_i is a Borel set of $(I \to \mathcal{M}_{r-1}(V))$. Hence we have

$$1_B \circ \hat{\mathbf{X}}^r = 1_{B_0} \circ \hat{\mathbf{X}}^r_0 \times \ldots \times 1_{B_T} \circ \hat{\mathbf{X}}^r_T.$$

Furthermore, since $\mathrm{Meas}(I \to \mathcal{M}_{r-1}(V))$ carries the Borel σ–algebra generated by the sets of the form

$$\begin{aligned} &e_U^{-1}(J) := \{\mu \in \mathrm{Meas}(I \to \mathcal{M}_{r-1}(V)) : e_U(\mu) := \mu(U) \in J\}, \\ &U \text{ Borel set in } (I \to \mathcal{M}_{r-1}(V)), \quad J \subseteq [0,1], \end{aligned}$$

we may use Dynkin's lemma again and assume that $B_i = e_{U_i}^{-1}(J_i)$ for some Borel set U_i in $(I \to \mathcal{M}_{r-1}(V))$ and some interval $J \subseteq [0,1]$. Now, using that $\hat{\mathbf{X}}^r_t = \mathbb{P}(\hat{\mathbf{X}}^{r-1} \in \cdot | \mathcal{F}_t)$, it holds that for all t,

$$1_{B_t} \circ \hat{\mathbf{X}}^r_t = 1_{J_t} \circ \mathbb{E}[1_{U_n} \circ \hat{\mathbf{X}}^{r-1} | \mathcal{F}_t].$$

By the induction hypothesis, we have

$$1_{U_n} \circ \hat{\mathbf{X}}^{r-1} = \lim_{k\to\infty} f_k^n(\mathbf{X}),$$

where every f_k^n is of rank at most $r-1$ and is uniformly bounded, so every $\mathbb{E}[f_k^n(\mathbf{X})|\mathcal{F}_t]$ is of rank at most r. Now we choose a sequence of uniformly bounded continuous functions $(\varphi_k)_{k\geq 1}$ (say, uniformly bounded by 1) which approximates $1_{J_0} \times \ldots \times 1_{J_I}$ pointwise, so that $1_B \circ \hat{\mathbf{X}}^r = \lim_{j\to\infty} \varphi_j(\hat{\mathbf{X}}_0^r, \ldots, \hat{\mathbf{X}}_T^r)$ a.s. (up to taking a subsequence if necessary). From the above observations we see that for each j,

$$\varphi_j(\hat{\mathbf{X}}_0^r, \ldots, \hat{\mathbf{X}}_T^r) = \lim_{k\to\infty} \varphi_j((\mathbb{E}[f_k^n(\mathbf{X})|\mathcal{F}_t])_{t\in I}),$$

where every $\varphi_j((\mathbb{E}[f_k^n(\mathbf{X})|\mathcal{F}_t])_{t\in I})$ is by definition an adapted functional of rank at most r. This shows that we can find a sequence of adapted functionals $(f_k)_{k\geq 1}$ of rank at most r, such that $1_B \circ \hat{\mathbf{X}}^r = \lim_{k\to\infty} f_k(\mathbf{X})$ in probability. □

Proof of Theorem 4.4.2. 1 $\implies$ 2. Using Lemma 4.4.3, it follows by an induction argument that $1_B \circ \hat{\mathbf{X}}^r = \lim_{k\to\infty} f_k(\mathbf{X})$ implies that $1_B \circ \hat{\mathbf{Y}}^r = \lim_{k\to\infty} f_k(\mathbf{Y})$. By (1), we have that $\mathbb{E}[f_k(\mathbf{X})] = \mathbb{E}[f_k(\mathbf{Y})]$ for all $k \geq 0$, so by the dominated convergence theorem

$$\mathbb{E}[1_B \circ \hat{\mathbf{X}}^r] = \mathbb{E}[1_B \circ \hat{\mathbf{Y}}^r] \text{ for any Borel set } B$$

i.e. $\text{Law}(\hat{\mathbf{X}}^r) = \text{Law}(\hat{\mathbf{Y}}^r)$.

2 $\implies$ 3. For a Polish space $\mathcal{X}$, let $\text{Meas}_a(\mathcal{X}) \subset \text{Meas}(\mathcal{X})$ be the set of Dirac measures on $\mathcal{X}$ $\{\delta_x : x \in \mathcal{X}\}$. Define $p : \text{Meas}_a(\mathcal{X}) \to \mathcal{X}$ by $p(\delta_x) := x$ and note that p is continuous with respect to the subspace topology on $\text{Meas}_a(\mathcal{X})$. Define

$$\pi_T : (I \to \mathcal{X}) \to \mathcal{X}, \quad \pi_T(x) = x_T \text{ for } x = (x_1, \ldots, x_T) \in (I \to \mathcal{X}),\ I = \{1, \ldots, T\}.$$

For $g : \mathcal{X} \to \mathcal{X}$ define $\text{id}_{\mathcal{X}} \oplus g : \mathcal{X} \to \mathcal{X}^2$, as $(\text{id}_{\mathcal{X}} \oplus g)(x) = (x, g(x))$. In what follows, although the underlying space $\mathcal{X}$ may vary from line to line, we will use the same notation as above for simplicity. Since $\hat{\mathbf{X}}_T^r = \mathbb{P}(\hat{\mathbf{X}}^{r-1} \in \cdot|\mathcal{F}_T)$, we can write

$$\hat{\mathbf{X}}_T^i = \delta_{\hat{\mathbf{X}}^{i-1}} \in \text{Meas}_a(I \to \text{Meas}_{r-1}).$$

For each r, define

$$g_r : I \to \text{Meas}_r,\ g_r = p \circ \pi_T.$$

Using that $g_r(\hat{\mathbf{X}}^r) = \hat{\mathbf{X}}^{r-1}$, it follows that for $r \geq 1$,

$$(\hat{\mathbf{X}}^r, \ldots, \hat{\mathbf{X}}^{r-s}) = (\text{id} \oplus g_{r-s+1}) \circ (\hat{\mathbf{X}}^r, \ldots, \hat{\mathbf{X}}^{r-s+1}),$$

where id is applied to $\mathcal{X} = \text{Meas}_r^I \times \cdots \times \text{Meas}_1^I$. Since, $\text{id} \oplus g_r$ is continuous we can iterate this composition to build a continuous function G, such that

$$(\hat{\mathbf{X}}^r, \ldots, \hat{\mathbf{X}}^0) = G(\hat{\mathbf{X}}^r).$$

As a result, for any bounded continuous function F defined on $\text{Meas}_r^I \times \ldots \times \text{Meas}_0^I$, we have

$$\mathbb{E}[F(\hat{\mathbf{X}}^r, \ldots, \hat{\mathbf{X}}^0)] = \mathbb{E}[F \circ G(\hat{\mathbf{X}}^r)],$$

Using 2 and denoting for brevity

$$E := (I \to \text{Meas}_r(\mathcal{X})) \times \cdots \times (I \to \text{Meas}_1(\mathcal{X}))$$

we deduce that

$$\begin{aligned}
\text{Law}(\hat{\mathbf{X}}^r) = \text{Law}(\hat{\mathbf{Y}}^r) &\Rightarrow \forall F \in C_b(E), \quad \mathbb{E}[F \circ G(\hat{\mathbf{X}}^r)] = \mathbb{E}[F \circ G(\hat{\mathbf{Y}}^r)] \\
&\Leftrightarrow \forall F \in C_b(E), \quad \mathbb{E}[F(\hat{\mathbf{X}}^r, \ldots, \hat{\mathbf{X}}^0)] = \mathbb{E}[F(\hat{\mathbf{Y}}^r, \ldots, \hat{\mathbf{Y}}^0)] \\
&\Leftrightarrow \text{Law}(\hat{\mathbf{X}}^0, \ldots, \hat{\mathbf{X}}^r) = \text{Law}(\hat{\mathbf{Y}}^0, \ldots \hat{\mathbf{Y}}^r).
\end{aligned}$$

3 $\implies$ 1. We prove by induction that for any $r \geq 0$, and $f \in \text{AF}_r$, there exists some bounded and Borel measurable $\tilde{f} : (I \to \text{Meas}_r(V)) \to \mathbb{R}$

$$f(\mathbf{X}) = \tilde{f}(\hat{\mathbf{X}}^r).$$

The case $r = 0$ is clear since $\hat{\mathbf{X}}^0 = X$ so we can take $\tilde{f} = f$, which is indeed bounded and Borel measurable.

For the induction step, assume the claim holds up to some $r - 1$, $r \geq 2$. Then given $f \in \text{AF}_{r-1}$, there exists a bounded Borel measurable function $\tilde{f}$ defined on $(I \to \text{Meas}_{r-1}(V))$ such that $f(\mathbf{X}) = \tilde{f}(\hat{\mathbf{X}}^{r-1})$. Now for every $t \in I$, $\mathbb{E}[f(\mathbf{X})|\mathcal{F}_t]$, is an element of AF_r, so by using that $f(\mathbf{X}) = \tilde{f}(\hat{\mathbf{X}}^{r-1})$, we get $\mathbb{E}[f(\mathbf{X})|\mathcal{F}_t] = \mathbb{E}[\tilde{f}(\hat{\mathbf{X}}^{r-1})|\mathcal{F}_t]$.

On the other hand, since by definition $\hat{\mathbf{X}}_t^r = \mathbb{P}(\hat{\mathbf{X}}^{r-1} \in \cdot|\mathcal{F}_t)$ is the regular conditional distribution of $\hat{\mathbf{X}}^{r-1}$ given $\mathcal{F}_t$, we also obtain that

$$\mathbb{E}[\tilde{f}(\hat{\mathbf{X}}^{r-1})|\mathcal{F}_t] = e_{\tilde{f}}(\pi_t \circ \hat{\mathbf{X}}^r),$$

where π_t is the t–th coordinate mapping such that $\pi_t \circ \hat{\mathbf{X}}^r = \hat{\mathbf{X}}_t^r$, and $e_{\tilde{f}}$ is the evaluation map defined on $\text{Meas}_r(V) = \text{Meas}(I \to \text{Meas}_{r-1}(V))$ such that $e_{\tilde{f}}(\mu) := \int \tilde{f} d\mu$. Since $\tilde{f}$ is bounded and measurable by [18, Corollary 7.29.1], $e_{\tilde{f}}$ is a bounded measurable function on $\text{Meas}_r(V)$.

In other words, we have now obtained that $\mathbb{E}[f(\mathbf{X})|\mathcal{F}_t] = g(\hat{\mathbf{X}}^r)$, where $g := e_{\tilde{f}} \circ \pi_t$ is a bounded measurable mapping defined on $\mathcal{X}_r$. This together with the fact that $\hat{\mathbf{X}}^{r-1}$ can be expressed as a Borel measurable function composition with $\hat{\mathbf{X}}^r$ (see the proof of (2) $\Rightarrow$ (3)) implies that all adapted functionals of rank at most r still satisfy the above claim, and completes the induction step.

2 $\iff$ 4. By Proposition 4.3.4, $\bar{\text{S}}_r$ is injective on $\text{Prob}_r(V)$ hence the equivalence follows immediately from Proposition 4.4.2 and the fact that $\text{Law}(\hat{\mathbf{X}}^r), \text{Law}(\hat{\mathbf{Y}}^r) \in \text{Prob}_{r+1}(V)$. □

The metrics d_r (cf. Definition 31) locally characterize the rank r extended weak topology $\hat{\tau}_r$.

Proposition 4.4.4. *Let V be a separable Banach space, $(\mathbf{X}^n)_{n\geq 0} \subset \mathcal{S}(V)$, $\mathbf{X} \in \mathcal{S}(V)$, and $r \geq 0$.*

1. *If $(\mathbf{X}^n)$ converges to $\mathbf{X}$ in $(\mathcal{S}(V), \hat{\tau}_r)$, then $d_r(\mathbf{X}^n, \mathbf{X}) \to 0$ as $n \to \infty$.*
2. *If $(\mathbf{X}^n)_{n\geq 0}$ is contained in a compact set of $(\mathcal{S}(V), \hat{\tau}_r)$, then $d_r(\mathbf{X}^n, \mathbf{X}) \to 0$ as $n \to \infty$ implies that $(\mathbf{X}^n)$ converges to $\mathbf{X}$ in $(\mathcal{S}(V), \hat{\tau}_r)$.*

Proof. 1: $\mathbf{X}^k$ converging to $\mathbf{X}$ in the rank r extended weak topology means that $\mathrm{Law}(\hat{\mathbf{X}}^{k,r})$ converges to $\mathrm{Law}(\hat{\mathbf{X}}^r)$. By Proposition 4.4.2, $\bar{\mathbb{E}}[\mathbf{X}_0^{r+1}] = \mathbb{E}[\mathrm{S} \circ \bar{\mathrm{S}}_r(\hat{\mathbf{X}}^r)]$, and by Proposition 4.3.4, $\mathrm{S} \circ \bar{\mathrm{S}}_r$ is a continuous and bounded function on $\mathcal{P}_r(V)$. The implication follows immediately.
2: By assumption, $(\mathbf{X}^k)_{k\geq 0}$ is contained in a compact set with respect to the rank r extended weak topology on $\mathcal{S}(V)$. Hence, there exists a $\mathbf{Y} \in \mathcal{S}(V)$ such that $\mathbf{X}^k$ converges to $\mathbf{Y}$ in the rank r extended weak topology. From the proof of claim 1 above, we have $d_r(\mathbf{X}^k, \mathbf{Y}) \to 0$. Hence $d_r(\mathbf{X}, \mathbf{Y}) = 0$, or equivalently, $\|\bar{\mathbb{E}}\bar{\mathbf{X}}_0^{r+1} - \bar{\mathbb{E}}\bar{\mathbf{Y}}_0^{r+1}\|_{r+1} = 0$. Now using Theorem 4.4.2 we obtain that $\mathrm{Law}(\hat{\mathbf{X}}^r) = \mathrm{Law}(\hat{\mathbf{Y}}^r)$.

□

We now relate d_r and the rank r extended weak topology with the adapted topology of rank r, τ_r (cf. Definition 22).

Proposition 4.4.5. *For a separable Banach space V, $\tau_r \subset \hat{\tau}_r$. That is, convergence of $(\mathbf{X}^n)$ in $(\mathcal{S}(V), \hat{\tau}_r)$ to $\mathbf{X}$ implies convergence of $(\mathbf{X}^n)$ to $\mathbf{X}$ in $(\mathcal{S}(V), \tau_r)$. Moreover, for processes evolving in a compact state space $\mathcal{K}$, $\mathcal{K} \subset V$, the converse holds. That is*

$$(\mathcal{S}(\mathcal{K}), \tau_r) = (\mathcal{S}(\mathcal{K}), \hat{\tau}_r).$$

Proof. First, it is easy to use an induction argument to show that for every $r \geq 0$, for every $f \in \mathrm{AF}_r$, there exists a bounded continuous function ρ_f defined on $I \to \mathcal{P}_r(V)$ such that $f(\mathbf{X}) = \rho_f(\hat{\mathbf{X}}^r)$ for all $\mathbf{X} \in \mathcal{S}(V)$. As a consequence, if $\mathbf{X}^k$ converges to $\mathbf{X}$ in the rank r extended weak topology; that is, $\mathrm{Law}(\hat{\mathbf{X}}^{k,r})$ converges weakly to $\mathrm{Law}(\hat{\mathbf{X}}^r)$ in $\mathcal{P}(I \to \mathcal{P}_r(V))$, then it indeed holds that for any $f \in \mathrm{AF}_r$

$$\lim_{k\to\infty} \mathbb{E}[\rho_f(\hat{\mathbf{X}}^{k,r})] = \mathbb{E}[\rho_f(\hat{\mathbf{X}}^r)]$$

which is equivalent to $\lim_{k\to\infty} \mathbb{E}[f(\hat{\mathbf{X}}^{k,r})] = \mathbb{E}[f(\hat{\mathbf{X}}^r)]$; i.e., $\mathbf{X}^k$ converges to $\mathbf{X}$ in the adapted topology of rank r. Also note that this result holds without assuming that

$(\mathbf{X}^k)_{k\geq 0}$ is contained in a compact set with respect to the rank r extended weak topology on $\mathcal{S}(V)$.
On the other hand, by the definition of adapted functionals (cf. Definition 19) one can easily verify that the class $\mathcal{A} := \{\rho_f : f \in \mathrm{AF}_r\}$ is a subalgebra in $C_b(I \to \mathcal{P}_r(\mathcal{K}); \mathbb{R})$. Moreover, using the proof of 1 $\implies$ 2 in Theorem 4.4.2 we can also prove that $\mathcal{A}$ separate points on $(I \to \mathcal{P}_r(\mathcal{K}))$. Therefore, since the space $(I \to \mathcal{P}_r(\mathcal{K}))$ is obviously compact (recall that $\mathrm{Prob}(\mathcal{K})$ is compact in the weak topology, and then by induction one can prove the compactness for all $\mathcal{P}_r(\mathcal{K})$), $\mathcal{A}$ is dense in $C_b(K; \mathbb{R})$ under the uniform topology by the Stone–Weierstrass theorem. Hence, for any given $\rho \in C_b(I \to \mathcal{P}_r(\mathcal{K}); \mathbb{R})$ and any $\varepsilon > 0$, we can pick a $\rho_f \in \mathcal{A}$ such that $\sup_{x \in (I \to \mathcal{P}_r(\mathcal{K}))} |\rho(x) - \rho_f(x)| \leq \varepsilon$, and deduce that

$$\begin{aligned} |\mathbb{E}[\rho(\hat{\mathbf{X}}^{k,r})] - \mathbb{E}[\rho(\hat{\mathbf{X}}^{r})]| &\leq |\mathbb{E}[\rho_f(\hat{\mathbf{X}}^{k,r})] - \mathbb{E}[\rho_f(\hat{\mathbf{X}}^{r})]| + 2\varepsilon \\ &= |\mathbb{E}[f(\hat{\mathbf{X}}^{k})] - \mathbb{E}[f(\hat{\mathbf{X}})]| + 2\varepsilon \to 0, \end{aligned}$$

where the last convergence holds as $\mathbf{X}^k$ converges to $\mathbf{X}$ in the adapted topology of rank r. □

Putting everything together, gives the following Theorem 4.4.3 which in turn implies Theorem 4.4.1.

Theorem 4.4.3. *Let V be a separable Banach space and $r \geq 0$. Then for $\mathbf{X}, \mathbf{Y} \in \mathcal{S}(V)$*

$$(\mathbb{E}[f(\mathbf{X})] = \mathbb{E}[f(\mathbf{Y})] \quad \forall f \in \mathrm{AF}_r) \text{ if and only if } \Phi_r(\mathbf{X}) = \Phi_r(\mathbf{Y}).$$

Moreover,

1. *the map Φ_r locally induces the topology τ_r,*
2. *the semimetric d_r locally metrizes the topology τ_r.*

If $\mathcal{K} \subset V$ is compact, then the above statements apply without localization, as stated in Theorem 4.4.1. in this case of a compact state space, one can also replace the robust signature in the definition of Φ_r with the classical signature.

Restricted to $r = 0$, we recover the fact that the expected signature can (locally) metrize weak convergence [46]; restricted to $r = 1$ this (locally) induces Aldous extended weak topology [3]. With $r = 1$ and $(\mathcal{F}_t)_{0\leq t\leq T}$ the natural filtration it adds another entry to the list in [6] of distances that induce the same topology, and therefore we obtain Corollary 4.4.1.

Remark 4.4.1. A natural question is that of quantitative guarantees and put simply, we do not have any non-trivial quantitative guarantees at this stage and we believe this is an interesting future research question. For example, for the causal Wasserstein distance, recent progress in [6, 10] shows that the value function

$$v : \mathbf{X} = (\Omega, (\mathcal{F}_t)_{t\in I}, \mathbb{P}, (X_t)_{t\in I}) \mapsto \inf_{\gamma} \mathbb{E}[L_\tau]$$

is even Lipschitz continuous with respect to the adapted Wasserstein metric on $\mathcal{S}$ but we suspect that in generality, Lipschitz continuity does not hold for our metric. Similarly, obtaining convergence rates for finite sample estimators, analogous to what is done in [7] for the causal Wasserstein distance, seems yet another promising research direction.

4.4.3 Tightness in Extended Weak topology: a Brief Discussion

For the case $V = \mathbb{R}^d$ one can obtain a tightness criterion for extended weak topology. First we note that in this case d_0 really metrizes the usual weak convergence on $\mathbb{R}^d$.

Proposition 4.4.6. *μ_n converges to μ weakly in* $\mathrm{Prob}(I \to \mathbb{R}^d)$ *if and only if* $\lim_{n\to\infty} d_0(\mu_n, \mu) = 0$.

The proof of this proposition is given in Appendix 4.C, see Proposition 4.C.2 and Corollary 4.C.3.
Now we consider $r = 1$; i.e, the Aldous' extended weak topology. Let $(\mu_t)_{t\in I}$ be a (discrete time) path in $(I \to \mathrm{Prob}(I \to \mathbb{R}^d))$ such that $\mu_t \in \mathrm{Prob}(I \to \mathbb{R}^d)$ for all $t \in I$. As before let S denote a normalized signature map on $I \to \mathbb{R}^d$. Then associated with $(\mu_t)_{t\in I}$ we obtain a $\mathbf{T}^1(\mathbb{R}^d)$–valued (discrete time) path $(\int \mathrm{S}(x)\mu_t(dx))_{t\in I}$. This mapping will be denoted by F, it is easy to see (by the Skorokhod representation theorem) that $F : (I \to \mathrm{Prob}(I \to \mathbb{R}^d)) \to (I \to \mathbf{T}^1(\mathbb{R}^d))$, $F((\mu_t)_{t\in I}) = (\int \mathrm{S}(x)\mu_t(dx))_{t\in I}$, is continuous.

Lemma 4.4.7. *A set* $\mathrm{U} \subset \mathcal{P}_2(\mathbb{R}^d)$ *is tight if and only if the set*

$$\{F_\sharp \mu : \mu \in \mathrm{U}\} \subset \mathcal{P}_1(\mathbf{T}^1(\mathbb{R}^d))$$

is tight, where $F_\sharp$ means the pushforward operation under F.

Proof. Let $\mathrm{Im}(F) \subset (I \to \mathbf{T}^1(\mathbb{R}^d))$ be the image of F, which is endowed with the subspace topology inherited from the Hilbert space $I \to \mathbf{T}^1(\mathbb{R}^d) = \mathbf{T}^1(\mathbb{R}^d)^I$. By Proposition 4.4.6 and the construction of d_0, we see that $F : (I \to \mathrm{Prob}(I \to \mathbb{R}^d)) \to \mathrm{Im}(F)$ is a homeomorphism. Hence the claim follows easily. □

Now let $\mathbf{X} = ((\Omega, \mathbb{P}, (\mathcal{F}_t)_{t\in I}), X) \in \mathcal{S}(\mathbb{R}^d)$, we know that $\hat{\mathbf{X}}^1 \in (I \to \mathrm{Prob}(I \to \mathbb{R}^d))$ and $\mathrm{Law}(\hat{\mathbf{X}}^1) \in \mathrm{Prob}_2(\mathbb{R}^d)$. Note also that $F(\hat{\mathbf{X}}^1) = (\mathbb{E}[\mathrm{S}(X)|\mathcal{F}_t])_{t\in I}$ is a $\mathbf{T}^1(\mathbb{R}^d)$–valued martingale, and thus $F_\sharp(\mathrm{Law}(\hat{\mathbf{X}}^1))$ is a martingale law on $I \to \mathbf{T}^1(\mathbb{R}^d)$. Hence, by Lemma 4.4.7 we obtain the following characterization of tightness set for the extended weak topology:

Theorem 4.4.4. *Let* $\mathrm{U} \subset \mathcal{S}(\mathbb{R}^d)$.

1. U *is tight in the Aldous' extended weak topology if and only if the collection of martingale laws*
$$\{\mathrm{Law}((\mathbb{E}[\mathrm{S}(X)|\mathcal{F}_t])_{t\in I}) : \mathbf{X} \in \mathrm{U}\}$$
is tight in $\mathrm{Prob}(I \to \mathbf{T}^1(\mathbb{R}^d))$.

2. *If in addition that all filtrations are natural and U is closed in the Aldous' extended weak topology, then the following are equivalent*
   - U *is compact in the Aldous' extended weak topology, Hellwig's information topology, and all other topologies mentioned in Corollary 4.4.1.*
   - U *satisfies the Eder's conditions ([66, Theorem 1.4]).*
   - *The collection of martingale laws* $\{\mathrm{Law}((\mathbb{E}[\mathrm{S}(X)|\mathcal{F}_t])_{t\in I}) : \mathbf{X} \in \mathrm{U}\}$ *is tight in* $\mathcal{P}_1(\mathbf{T}^1(\mathbb{R}^d))$.

Remark 4.4.2.

1. The assumption that $V = \mathbb{R}^d$ cannot be removed because we need the local compactness of $(I \to \mathbb{R}^d)$ to ensure Proposition 4.4.6. This observation also prevent us from extending above results to higher rank extended weak topologies as, for example, the space $\mathrm{Prob}(I \to \mathbb{R}^d)$ is not locally compact in general. On the other hand, if $V = \mathbb{R}^d$, we know from [10, Theorem 1.7] that the tightness of subsets in $(\mathcal{S}(V), \hat{\tau}_r)$ is equivalent to the tightness of subsets in $(\mathcal{S}(V), \hat{\tau}_0)$ for any $r \geq 0$. Hence, in view of Proposition 4.4.6, in this case we have the metric d_r metrizes the topology $\hat{\tau}_r$ on $\mathcal{S}(V)$ for all $r \geq 0$ (not only locally).

2. Theorem 4.4.4 complements Eder's tightness theorem [66, Theorem 1.4]. In particular, we highlight that the expected signature map transforms the tightness of laws on a measure space into tightness of martingale laws on a Hilbert space. This allows to use tools from martingale theory and the Hilbert space to study the extended weak topology, e.g. to define the Fourier transform (characteristic

functions) for the law of prediction processes. Further, it suggests a concrete numerical way to check the tightness in extended weak topology by formulating it in the tensor algebra $\mathbf{T}^1(\mathbb{R}^d)$;

4.5 Algorithms and Experiments

In this Section we apply dynamic programming principles to derive algorithms that efficiently compute $\Phi_r(\mathbf{X})$ when the process $\mathbf{X}$ is a Markov chain.

We will assume that $\mathbf{X}$ is *nowhere recombining*, that is that its value at time t uniquely determines $X_1, \dots, X_{t-1}$. This is satisfied for a large class of Markov processes and by echoing the usual remark that every process is Markovian by lifting it to a process in larger state space[9], any process $\mathbf{X}$ can be made to satisfy the assumptions of this section by considering instead the process $Z_t = (X_1, \dots, X_t)$.

In the construction of $\Phi_r(\mathbf{X})$, the process $\mathbf{X}$ is lifted to a process that evolves in the algebra $\mathbf{T}^r(V)$ and dynamic programming naturally applies there as the following lemma shows.

Lemma 4.5.1. *Let A be an algebra and $\mathbf{Z} \in \mathcal{S}(A)$ be a nowhere recombining Markov chain with finite support. The function*

$$u_t(a) := \mathbb{E}[Z_{t+1} \cdots Z_T | Z_t = a]$$

satisfies for every $t = 0, 1, \dots, T$ and $a \in A$ the recursion

$$u_t(a) = \sum_{b \in A} b\mathbb{P}(Z_{t+1} = b | Z_t = a) u_{t+1}(b)$$

Proof. This follows immediately from

$$u_t(a) = \mathbb{E}[Z_{t+1} \cdots Z_T | Z_t = a] = \mathbb{E}[Z_{t+1} u_{t+1}(Z_{t+1}) | Z_t = a].$$

□

If $\mathbf{X} \in \mathcal{S}(V)$ is a nowhere recombining Markov chain, then the process $\mathbf{Z} \in \mathcal{S}(\mathbf{T}^1(V))$ defined as

$$Z_t := \exp(\Delta_t X) \text{ with } \Delta_t X := (X_{t+1} - X_t, 1)$$

[9]One may consider the path valued random variable $\mathbf{Y}$ such that the values of Y_t are $(X_0, \dots, X_t)$. This "lifting" is trivial for processes with small state spaces.

where $\exp : V \to \mathbf{T}^1(V)$ denotes the tensor exponential, can also be lifted to a nowhere recombining Markov chain that takes values in the algebra $\mathbf{T}^1(V)$. Recall that the signature $\mathrm{S}(x)$ of a path $x : I \to V$ is defined as

$$\mathrm{S}(x) = \exp(\Delta_0 x) \exp(\Delta_1 x) \cdots \exp(\Delta_T x).$$

Hence, Lemma 4.5.1 hints at an efficient way to compute $\Phi_0(\mathbf{X}) \equiv \mathbb{E}[\mathrm{S}(X)] = u_0(X_0)$ since the value function u satisfies the recursion:

$$u_t(x) = \sum_{y \in V} \exp(y - x) \mathbb{P}(X_{t+1} = y | X_t = x) u_{t+1}(y).$$

Algorithm 1 formulates this in pseudo-code by representing a Markov chain $\mathbf{X}$ as tree: vertices are labelled by the attainable states V of the Markov chain $\mathbf{X}$; the process starts at time $t = 0$ at a root vertex $r = X_0$; if the Markov chain at time t has value a then we denote by *a.children* the set of attainable states (vertices) at time $t + 1$; the transition probability between two states a and b is denoted by $p(a, b)$.

Algorithm 1 Pseudo-code for $\Phi_0(\mathbf{X})$

Input: A Markov chain $\mathbf{X}$ represented as a rooted tree with root r
procedure $\mathrm{ExpSig}_0(a)$
    if *a.children* **is** empty **then**
        return 1
    sum $\leftarrow 0$
    for b **in** *a.children* **do**
        sum $\leftarrow$ sum $+ p(a, b) \cdot \exp\{b - a\} \cdot \mathrm{ExpSig}_0(b)$
    return sum
Output: $\mathrm{ExpSig}_0(r) = \Phi_0(\mathbf{X})$

Algorithm 2 Pseudo-code for $\Phi_1(\mathbf{X})$

Input: A Markov chain $\mathbf{X}$ represented as a rooted tree with root r, $s(a)$ denotes the signature of the sample path of $\mathbf{X}$ that ends at a.
procedure $\mathrm{ExpSig}_1(a)$
    $a_1 \leftarrow 0$
    $a_2 \leftarrow 0$
    for b **in** *a.children* **do**
        $b_1, b_2 \leftarrow \mathrm{ExpSig}_1(b)$
        $a_1 \leftarrow a_1 + p(a, b) * \exp\{b - a\} * b_1$
    for b **in** *a.children* **do**
        $b_1, b_2 \leftarrow \mathrm{ExpSig}_1(b)$
        $a_2 \leftarrow a_2 + p(a, b) * \exp\{s(b) * b_1 - s(a) * a_1\} * b_2$
    return a_1, a_2
Output: $\mathrm{ExpSig}_1(r) = \Phi_1(\mathbf{X})$

Lemma 4.5.2. *Let $\mathbf{X} \in \mathcal{S}(U)$ be a nowhere recombining Markov chain. If the rooted tree that represents $\mathbf{X}$ has at most $N+1$ vertices and depth d (that is, $\mathbf{X}$ terminates at time d) then Algorithm 1 computes $\Phi_1(\mathbf{X})$ with complexity*

$$O\Big(tN\Big) \text{ in time and } O\Big(ds\Big) \text{ in space,}$$

where t, s are the time and space costs of computing and storing one call of $\exp\Big(\,\cdot\,\Big)$ to the desired accuracy.

Proof. Note that the bounds are clearly true if the tree has a a root with N children that are all leaves as the recursion will visit each child once and needs to store the return value as well as the execution stack of depth 1. Recursively, if the root has n children, each of which is the root of a sub-tree with $n_i + 1$ vertices and depth d_i for $i = 1, \dots, n$. Then the recursion visits each sub-tree once and adds the results of each sub-tree to the return value. The time and space complexities of the recursive call on the i:th child are $O(tn_i)$ and $O(sd_i)$ respectively, hence the total time complexity is

$$O(\sum_{i=1}^{n} tn_i) = O(tN).$$

The space complexity is the maximum amount of space needed for the recursive call, plus the extra space for storing the value at the root and the execution stack, hence the total space complexity is

$$O(1 + s + \max_{1 \le i \le n}(sd_i)) = O(1 + s(d+1)) = O(ds),$$

proving the assertion. □

The computation of $\Phi_r(\mathbf{X})$ for $r > 1$ follows along the same lines, but since the notation gets increasingly cumbersome as r increases we only spell out the case $r = 2$ in detail; the cases $r \ge 3$ follow analogous. We now apply Lemma 4.5.1 with $A := \mathbf{T}^2(V)$ and $Z_t = \exp(\Delta_t \bar{\mathbf{X}}^1)$ the function $v_t^2(x) := \mathbb{E}[Z_{t+1} \cdots Z_T | Z_t = x]$ satisfies the recursion

$$v_t^2(x) = \sum_{y \in V} \mathbb{P}(X_{t+1} = y | X_t = x) \exp\Big\{\mathbb{E}[\mathrm{S}(X)|X_{t+1} = y] - \mathbb{E}[\mathrm{S}(X)|X_t = x]\Big\} v_{t+1}^2(y).$$

This recursion is more involved as it requires two evaluations of $\mathbb{E}[\mathrm{S}(X)|X_t]$ at every step. However, if we assume that the process X is nowhere recombining we can rewrite this as the following system

$$v_t^2(x) = \sum_{y \in V} \mathbb{P}(X_{t+1} = y | X_t = x) \exp\Big\{\,\mathrm{S}(X|X_t = y) v_{t+1}^1(y) - \mathrm{S}(X|X_{t-1} = x) v_t^1(x)\Big\} v_{t+1}^2(y)$$
$$v_t^1(x) = \sum_{y \in V} \mathbb{P}(X_{t+1} = y | X_t = x) \exp\Big\{y - x\Big\} v_{t+1}^1(y).$$

where $\mathrm{S}(X|X_t = y)$ denotes the signature of the path $X_0, \dots, X_t$ such that $X_t = y$, which is well defined since X is nowhere recombining by assumption. Note that v^1 is the same function as the one defined in Equation 4.5. Unfortunately $v_t^2(x)$ depends on both v_t^1 and v_{t+1}^1 and since multiplication in $\mathbf{T}^2(V)$ is non-commutative there is no way to separate the two dependencies. Because of this $v_t^1(x)$ needs to be computed before $v_t^2(x)$ and the recursion is best solved using a Dynamic Programming approach, or by caching the relevant values at every function call. This approach is outlined in Algorithm 2.

Algorithm 2 is more involved than Algorithm 1 as it requires the computation of $\mathrm{S}(x)$ at every recursive call which has time complexity $O(dt)$ where t is the time costs of computing one call of $\exp_1\big(\big)$ to the desired accuracy, and d is the depth of x. This can be remedied by memoising the values of $\mathrm{S}(x)$ once computed, which brings the time complexity down to $O(t)$ but takes up more space.

Lemma 4.5.3. *Let T, S be the time and space costs of computing and storing one call of $\exp : \mathbf{T}^1(V) \to \mathbf{T}^2(V)$ to the desired accuracy, and t, s be the time and space costs of one call of $\exp : V \to \mathbf{T}^1(V)$. Assume that the tree has exactly $N+1$, depth d and maximal degree M. Then the function $ExpSig_2$ can be implemented to be*

$$O\Big((t+T)N\Big) \text{ in time and } O\Big(d(MS+s)\Big) \text{ in space.}$$

Proof. The same arguments made in the proof of Lemma 4.5.2 applies here too. If one caches $\mathrm{S}(X|X_t = a)$ at every node one needs to store at most d values of $\mathrm{S}(\cdot)$ and the cost of computing $\exp\Big\{\mathrm{S}(X|X_t = b)v_{t+1}^1(b) - \mathrm{S}(X|X_{t-1} = a)v_t^1(a)\Big\}$ is always $O(T+t)$. By caching values of v_{t+1} in the first pass the maximum amount of space needed for each pass is MS, since nodes are visited at most once the maximum amount of memory needed is dMS. □

Remark 4.5.1 (Representing higher rank tensor algebras). In order to implement any of the computations outlined above, one first needs to be able to represent the relevant algebras. $\mathbf{T}^1(V)$ is well known to be a graded algebra over V, but $\mathbf{T}^2(V)$ has a more complicated multi-grading, and higher dimensional components which make computations trickier. See Appendix 4.B for a more thorough discussion of how to write down the gradings, and the dimensions of the spaces involved, but note that it is always to write down (formal) gradings $\mathbf{T}^r(V) = \prod_{k\geq 0} \mathbf{T}^r(V)_k$, where if V has

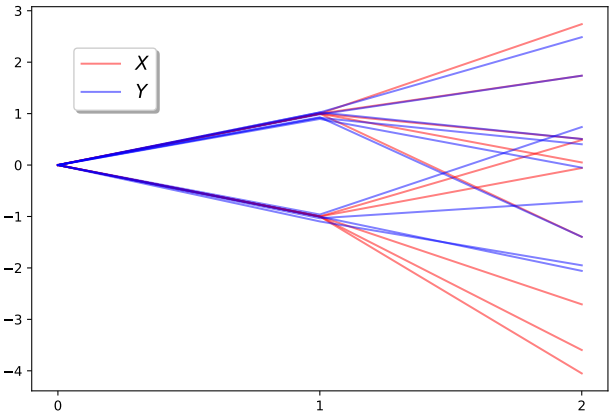


Figure 4.3: The sample paths of μ are plotted above for $\varepsilon = 0.05$. They are marked as either $\mathbf{X}$ in red, or $\mathbf{Y}$ in blue depending on if the sample comes from $\mathbf{X}^c$ or $\mathbf{Y}^c$.

dimension d, then

$$\begin{aligned}
&\dim.\mathbf{T}^1(V)_k = (d+1)^k, \\
&\dim.\mathbf{T}^2(V)_0 = 1, \quad \dim.\mathbf{T}^2(V)_1 = d+1, \\
&\dim.\mathbf{T}^2(V)_k = (2d+3)\dim.\mathbf{T}^2(V)_{k-1} - (d+1)\dim.\mathbf{T}^2(V)_{k-2}.
\end{aligned}$$

4.5.1 Experiment: Model Space and Linear Separability

Expected signatures (Φ_0 in our notation) are currently finding applications in machine learning. One of their attractive properties is that they provide a hierarchical description of the law of a stochastic process; in the terminology of statistical learning the signature map is a so-called "universal and characteristic" feature map for paths, see Appendix 4.C . However, expected signatures metrize weak convergence and hence completely ignore the filtration.

We now use the algorithms from the previous section to demonstrate on a simple numerical toy example that the geometry of the feature space of Φ_0 is too simple in the sense that it fails to separate models with different filtrations. In contrast, the feature space of Φ_1 is large enough to allow for a a linear separation.

Example 4.5.1 (Mixtures of Adapted Processes)**.** Define for every $c \in \mathbb{R}$ two processes $\mathbf{X}^c, \mathbf{Y}^c \in \mathcal{S}$ as

$$\begin{array}{lllll}
\mathbf{X}^c: & X_0 = 0, & X_1 = N_1, & & X_2 = c + N_2, \\
\mathbf{Y}^c: & Y_0 = 0, & Y_1 = \sqrt{1-\varepsilon^2}M_1 + \varepsilon c, & & Y_2 = c + M_2.
\end{array}$$

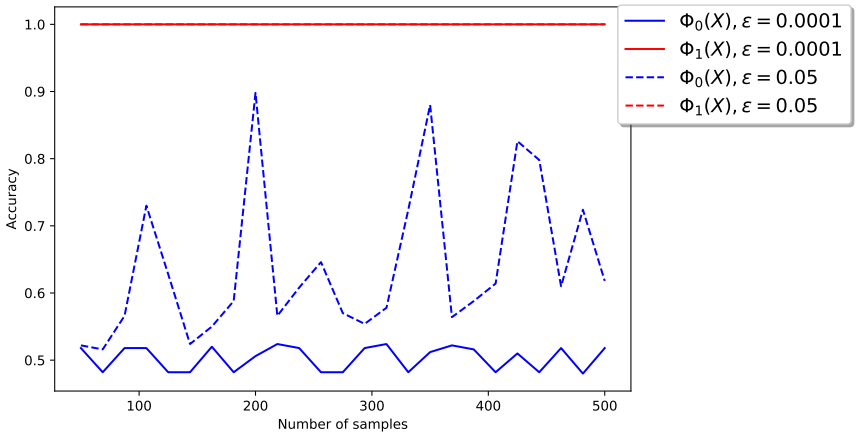


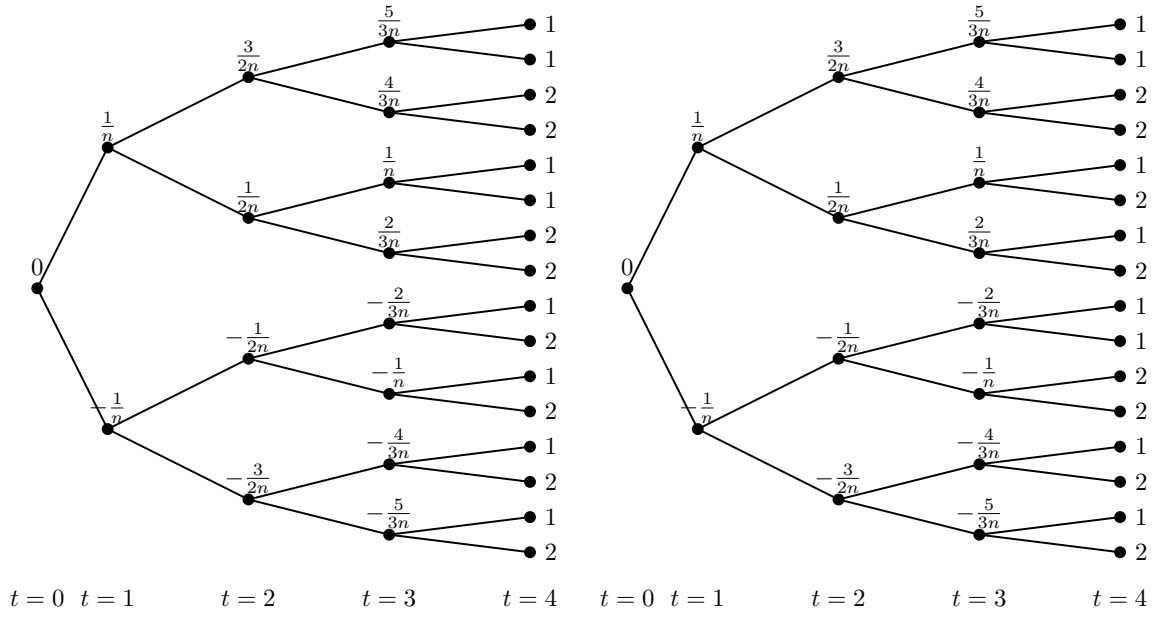
Figure 4.4: The accuracies of the linear classifier trained on $\Phi_0(\mathbf{X})$ and $\Phi_1(\mathbf{X})$ is plotted against the number of samples used in blue and red respectively. Solid lines are used for $\varepsilon = 10^{-4}$ and dashed lines for $\varepsilon = 5 \times 10^{-2}$.

where M_1, M_2, N_1, N_2 are independent Binomial random variables and $\varepsilon > 0$ is fixed. $\mathbf{X}^c$ and $\mathbf{Y}^c$ are both equipped with their natural filtrations. Note that for a fixed c and small ε, $\text{Law}(\mathbf{X}^c) \approx \text{Law}(\mathbf{Y}^c)$, analogously to Example 4.1.1 resp. Figure 4.1. $\mathcal{S}(\mathbb{R})$ is equipped with a probability measure μ as follows: a sample from μ consists of sampling a $C \sim N(0, 1)$ and then selecting with probability 0.5 the process X^C and with probability 0.5 the process Y^C. Figure 4.3 shows for each sample from μ (an adapted process) one sample trajectory from this process.

We ran the following experiment: We sampled 1000 processes from μ and labelled the processes corresponding to whether $\mathbf{X}^c$ or $\mathbf{Y}^c$ was sampled. We then computed Φ_0 and Φ_1 for each sample truncated at level 6 and level 3 respectively[10] and normalized the features. This was then split into a training set and test set – both of size 500 – and for $50 \leq m \leq 500$ a Support Vector Machine classifier [99] was trained on m data points from the training set with Φ_0 resp. Φ_1 as feature map.

Figure 4.4 shows the accuracies of the resulting classifiers on the test set. Observe that for small values of ε, the classifier on Φ_0 is essentially guessing, and even for larger values it does not converge well, the classifier on Φ_1 converges immediately however, which is to be expected as Φ_1 is able to separate $\mathbf{X}^c$ and $\mathbf{Y}^c$ independently of the value of ε.

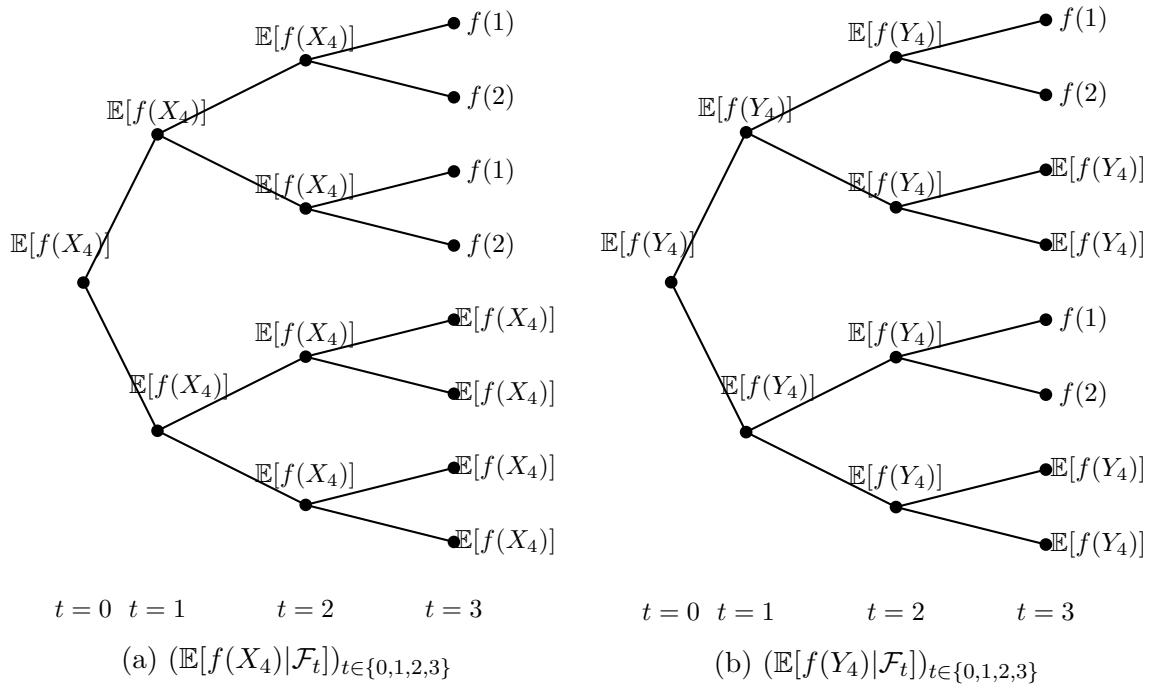
[10]This corresponds to 127 coordinates for Φ_0 resp. 76 coordinates for Φ_1, hence is in favour of Φ_0.



(a) $(X_t)_{t\in\{0,1,2,3,4\}}$ (b) $(Y_t)_{t\in\{0,1,2,3,4\}}$

Figure 4.5: Two processes X and Y that converge weakly to the same process and such that the difference between their prediction processes converges weakly to zero, but does not converge in the rank 2 adapted topology. Before $t = 4$ they move very little, with steps of order $1/n$ and at $t = 4$ they jump to either 1 or 2 with equal probability. As $n \to \infty$ they both converge weakly to the process that stay at 0 until $t = 4$ when it jumps to either 1 or 2.

We emphasize that although this is a toy example, it demonstrates how the expected signature can fail to pick up essential properties of a model and that the higher rank expected signatures provide additional features that linearise complex dependencies between law and filtration by representing them on multi-graded spaces of tensors.



(a) $(\mathbb{E}[f(X_4)|\mathcal{F}_t])_{t\in\{0,1,2,3\}}$ (b) $(\mathbb{E}[f(Y_4)|\mathcal{F}_t])_{t\in\{0,1,2,3\}}$

Figure 4.6: For any fixed $f \in C_b(\mathbb{R})$, the processes $t \mapsto \mathbb{E}[f(X_4)|\mathcal{F}_t]$ and $t \mapsto \mathbb{E}[f(Y_4)|\mathcal{F}_t]$ have the same distribution, so $\hat{X} - \hat{Y}$ goes to 0 as $n \to \infty$. The rank 2 prediction processes are not the same however

4.A Details for Example 4.1.2

Consider the Probability space $\Omega = \{1, \dots, 16\}$ equipped with the counting measure and the filtration

$$\begin{aligned}
\mathcal{F}_0 &= \{\Omega, \varnothing\},\\
\mathcal{F}_1 &= \sigma\langle\{1,\dots,8\},\{9,\dots,16\}\rangle,\\
\mathcal{F}_2 &= \sigma\langle\{1,2,3,4\},\{5,6,7,8\},\{9,10,11,12\},\{13,14,15,16\}\rangle,\\
\mathcal{F}_3 &= \sigma\langle\{1,2\},\{3,4\},\{5,6\},\{7,8\},\{9,10\},\{11,12\}\{13,14\},\{15,16\}\rangle,\\
\mathcal{F}_4 &= 2^{\Omega}.
\end{aligned}$$

Define the two processes

$$X_0 = X_1 = X_2 = X_3 = 0, X_4 : \begin{cases} 1,2,5,6,9,11,13,15 \mapsto 1 \\ 3,4,7,8,10,12,14,16 \mapsto 2, \end{cases}$$

$$Y_0 = Y_1 = Y_2 = Y_3 = 0, Y_4 : \begin{cases} 1,2,5,7,9,10,13,15 \mapsto 1 \\ 3,4,6,8,11,12,14,16 \mapsto 2, \end{cases} .$$

If the above construction looks unnatural the reader is also invited to think of the filtration as being the natural filtration associated to the processes and that instead

of staying at 0 until time 4, they move with step size of order $1/n$ in such a way to generate $\mathcal{F}$, as in Figure 4.5 and 4.6. Clearly the image measure of X and Y are the same, so $\mathbb{E}f(X) = \mathbb{E}f(Y)$ for any $f \in \mathbb{R}^{\{0,1,2\}}$. Moreover:

$$\mathbb{E}[f(X_4)|\mathcal{F}_0] = \mathbb{E}[f(X_4)|\mathcal{F}_1] = \mathbb{E}[f(X_4)|\mathcal{F}_2] = \frac{1}{2}\Big(f(1) + f(2)\Big),$$

$$\mathbb{E}[f(X_4)|\mathcal{F}_3] : \begin{cases} \{1,2\}, \{5,6\} \mapsto f(1) \\ \{3,4\}, \{7,8\} \mapsto f(2), \\ \{9,10\}, \{11,12\}, \{13,14\}, \{15,16\} \mapsto \frac{1}{2}\Big(f(1) + f(2)\Big), \end{cases}$$

$$\mathbb{E}[f(Y_4)|\mathcal{F}_0] = \mathbb{E}[f(Y_4)|\mathcal{F}_1] = \mathbb{E}[f(Y_4)|\mathcal{F}_2] = \frac{1}{2}\Big(f(1) + f(2)\Big),$$

$$\mathbb{E}[f(Y_4)|\mathcal{F}_3] : \begin{cases} \{1,2\}, \{9,10\} \mapsto f(1) \\ \{3,4\}, \{11,12\} \mapsto f(2), \\ \{5,6\}, \{7,8\}, \{13,14\}, \{15,16\} \mapsto \frac{1}{2}\Big(f(1) + f(2)\Big), \end{cases}$$

since the image measure of the above processes are the same, $\mathbb{E}[g(\mathbb{E}[f(X)|\mathcal{F}])] = \mathbb{E}[g(\mathbb{E}[f(Y)|\mathcal{F}])]$ for any $f, g \in \mathbb{R}^{\{0,1,2\}}$ and therefore they have the same prediction process. However, it can bee seen that

$$\mathbb{E}[\mathbb{E}[X_4|\mathcal{F}_3]^2|\mathcal{F}_1] : \begin{cases} \{1,2,3,4,5,6,7,8\} \mapsto \frac{5}{2} \\ \{9,10,11,12,13,14,15,16\} \mapsto \frac{9}{4}, \end{cases}$$

$$\mathbb{E}[\mathbb{E}[Y_4|\mathcal{F}_3]^2|\mathcal{F}_1] = \frac{19}{8}.$$

Hence the information structure in these processes are different, but this can't be seen by their prediction processes alone.

4.B Higher rank tensor algebras and their norms

If V is a Banach space with norm $\|\cdot\|$, then we want to equip $V^{\otimes m}$ with a norm for every $m \geq 1$. In the general case some care is needed and we assume that all norms on tensor products are *admissible* as defined below.

Definition 33. We say that $\|\cdot\|$ is an admissible norm on $(V^{\otimes m})_{m\geq 1}$, if:

1. For any permutation $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$

$$\|v_1 \otimes \cdots \otimes v_n\| = \|v_{\sigma(1)} \otimes \cdots \otimes v_{\sigma(n)}\|.$$

2. For $v \in V^{\otimes n}, w \in V^{\otimes m}$ it holds that

$$\|v \otimes w\| \leq \|v\| \cdot \|w\|.$$

Both projective, and injective norms are admissible. See [155] for more details.

Recall that if V is a vector space, then $\mathbf{ut}(V) := \bigoplus_{m\geq 0} V^{\otimes m}$ is the tensor algebra over V, and $\mathbf{ut}(V) - 1 := \bigoplus_{m\geq 1} V^{\otimes m}$ is the non-unital tensor algebra over V. The higher order tensor algebras are defined inductively as follows:

Definition 34. Let V be a normed space. Define the spaces

$$\mathbf{ut}^0(V) = V, \quad \mathbf{ut}^r(V) = \bigoplus_{m\geq 0} \Big(\mathbf{ut}^{r-1}(V) - 1\Big)^{\otimes m}, \qquad r \geq 1.$$
$$\mathbf{t}^0(V) = V, \quad \mathbf{t}^r(V) = \bigoplus_{m\geq 0} \Big(\mathbb{R} \oplus \mathbf{t}^{r-1}(V) - 1\Big)^{\otimes m}, \qquad r \geq 1.$$

A module A is said to be multi-graded if there is a monoid M such that

$$A = \bigoplus_{m\in M} A_m$$

and $A_m A_n \subseteq A_{mn}$. In order to describe the multi-grading of $\mathbf{ut}^r(V)$ and $\mathbf{t}^r(V)$ we will use the following lemma. We use the notation $\mathcal{F}\Big[\,\cdot\,\Big]$ for the free algebra generated by $\cdot$ and $\mathcal{M}\Big[\,\cdot\,\Big]$ for the free monoid generated by $\cdot$. The following follows from the definitions of $\mathcal{F}$ and $\mathcal{M}$ and is recorded here as a lemma.

Lemma 4.B.1. *Let $A_1, \ldots, A_n$ be multi-graded modules with respective multi-gradings $M_1, \ldots, M_n$. Then $\mathcal{F}\Big[A_1, \ldots, A_n\Big]$ is multi-graded by $\mathcal{M}\Big[M_1, \ldots, M_n\Big]$.*

Using the notation $\otimes_{(r)}$ for the tensor product on $\mathbf{ut}^{r-1}(V)$, we note that by the above lemma $\Big(\mathbf{ut}^r(V), +, \otimes_{(r)}\Big)$ is a multi-graded algebra over V. By recursively defining $\mathrm{Seq}^r := \mathrm{Seq}(\mathrm{Seq}^{r-1})$ with the convention that $\mathrm{Seq}^0 = \{\varnothing\}$. We may write down the multi-grading for $\mathbf{ut}^r(V)$ as follows

$$\mathbf{ut}^1(V)_{\mathbf{k}} := V^{\otimes_{(1)}\mathbf{k}}, \quad \mathbf{ut}^r(V) = \bigoplus_{\mathbf{k}\in \mathrm{Seq}^r} \mathbf{ut}^r(V)_{\mathbf{k}},$$

where $\mathbf{ut}^r(V)_{\mathbf{k}} := \mathbf{ut}^{r-1}(V)_{\mathbf{k}_1} \otimes_{(r)} \cdots \otimes_{(r)} \mathbf{ut}^{r-1}(V)_{\mathbf{k}_l}$ for $\mathbf{k} = \mathbf{k}_1 \cdots \mathbf{k}_l \in \mathrm{Seq}^r$.

We also use the following recursive definition of the degree for a multi-index

$$\deg.\mathbf{k} = \deg.\mathbf{k}_1 + \cdots + \deg.\mathbf{k}_l, \text{ for } \mathbf{k} = \mathbf{k}_1 \cdots \mathbf{k}_l \in \mathrm{Seq}^r, \quad \deg.\varnothing = 1.$$

Which allows us to write down a grading for $\mathbf{ut}^r(V)$ as

$$\mathbf{ut}^r(V) = \bigoplus_{k\geq 0} \Big(\bigoplus_{\substack{\mathbf{k}\in \mathrm{Seq}^r, \\ \deg.\mathbf{k}=k}} \mathbf{ut}^r(V)_{\mathbf{k}} \Big). \tag{4.10}$$

See [64, Section 3] for more on $\mathbf{ut}^1(V)$ and $\mathbf{ut}^2(V)$.

Example 4.B.1.

- $\mathbf{ut}^1(V)$ is the standard tensor algebra over V and is graded over $\mathrm{Seq}^1 \simeq \mathbb{N}$ by

$$\mathbf{ut}^1(V) = 1 + \bigoplus_{n \geq 1} V^{\otimes_{(1)} n}.$$

- $\mathbf{ut}^2(V)$ is graded over sequences in $\mathbb{N}$ by

$$\mathbf{ut}^2(V) = 1 + \bigoplus_{n_1, \ldots, n_k \geq 1} V^{\otimes_{(1)} n_1} \otimes_{(2)} \cdots \otimes_{(2)} V^{\otimes_{(1)} n_k}.$$

- $\mathbf{ut}^3(V)$ is graded over matrices in $\mathbb{N}$ by

$$\mathbf{ut}^3(V) = 1 + \bigoplus_{\substack{n_1^1, \ldots, n_{k_1}^1 \geq 1 \\ \vdots \\ n_1^{k_2}, \ldots, n_{k_1}^{k_2} \geq 1}} \left(V^{\otimes_{(1)} n_1^1} \otimes_{(2)} \cdots \otimes_{(2)} V^{\otimes_{(1)} n_{k_1}^1} \right) \otimes_{(3)} \cdots \otimes_{(3)} \left(V^{\otimes_{(1)} n_1^{k_2}} \otimes_{(2)} \cdots \otimes_{(2)} V^{\otimes_{(1)}}$$

Remark 4.B.1. For any ring R and R module M the above construction (disregarding the norm) yields a sequence of R algebras

$$M \subseteq \mathbf{ut}^1(M) \subseteq \mathbf{ut}^2(M) \subseteq \cdots.$$

This sequence is characterized by the following universal property which follows from the universal property of the tensor algebra:

For any R module N, $r \geq 0$ and R-module homomorphism $\varphi : \mathbf{ut}^r(M) \to N$ there exists a unique R-algebra homomorphism $\Phi : \mathbf{ut}^{r+1}(M) \to N$ such that $\varphi = \Phi \circ \iota$ where ι is the inclusion map $\mathbf{ut}^r(M) \hookrightarrow \mathbf{ut}^{r+1}(M)$.

Example 4.B.2. For a concrete example of this, if V is some vector space over $\mathbb{R}$ and X is a bounded random variable on V, then its associated moment map is the linear map

$$\mu_X : \mathbf{ut}^1(V) \to \mathbb{R}, \quad \mu_X(e_{i_1} \cdots e_{i_k}) = \mathbb{E}(X_{i_1} \cdots X_{i_k})$$

which induces the algebra homomorphism $\mu_X^\star : \mathbf{ut}^2(V) \to \mathbb{R}$. The reader familiar with *cumulants* might note that the cumulants of X can then be described as linear functions of $\mu_X^\star$. For example

$$\kappa(X_1, X_2, X_3) = \mathbb{E}(X_1X_2X_3) - \mathbb{E}(X_1)\mathbb{E}(X_2X_3) - \mathbb{E}(X_2)\mathbb{E}(X_1X_3) - \mathbb{E}(X_3)\mathbb{E}(X_1X_2) + 2\mathbb{E}(X_1)\mathbb{E}(X_2$$
$$= \mu_X^\star(e_1e_2e_3) - \mu_X^\star(e_1 \otimes_{(2)} e_2e_3) - \mu_X^\star(e_2 \otimes_{(2)} e_1e_3) - \mu_X^\star(e_3 \otimes_{(2)} e_1e_2) + 2\mu_X^\star(e_1 \otimes_{(2)} e_2 \otimes_{(2)} e_3)$$

4.B.1 The multi-grading of $\mathbf{t}^r(V)$

Recall that the signature map, Definition 25, takes paths on a vector space V as input and maps them into $\mathbf{T}^1(V)$ which is isomorphic to the completion of $\mathbf{ut}^1(V \oplus \mathbb{R})$, so that in order to represent the signature of a path in V it is enough to represent elements of $\mathbf{ut}^1(V)$ for arbitrary finite dimensional V.

In the rank two case we are not so lucky however, as $\mathbf{t}^2(V) = \mathbf{t}^1(\mathbf{t}^1(V)) \simeq \mathbf{ut}^1(\mathbf{ut}^1(V \oplus \mathbb{R}) \oplus \mathbb{R})$ which is not isomorphic to a rank 2 tensor algebra over any finite dimensional space. By definition, $\mathbf{t}^2(V)$ is the free algebra generated by $\mathbf{t}^1(V)$ and one indeterminate, so by Lemma 4.B.1 it is multi-graded by $\mathcal{M}(\mathcal{M}^1, \mathcal{M}^0)$. We may write:

$$\mathbf{t}^2(V) = \bigoplus_{m_1,\ldots,m_n \in \mathcal{M}^1,\mathcal{M}^0} \mathbf{t}^2(V)_{m_1} \otimes_{(2)} \cdots \otimes_{(2)} \mathbf{t}^2(V)_{m_n}$$

where $\mathbf{t}^2(V)_m = \mathbf{ut}^1(V)_m$ if $m \in \mathcal{M}^1$ and $\mathbf{t}^2(V)_m \simeq \mathbb{R}$ if $m \in \mathcal{M}^0$. This multi-grading also allows us to write down a grading for $\mathbf{t}^2(V)$ like in Equation (4.10) where the degree is defined in the natural way compatible with the degrees on $\mathcal{M}^1, \mathcal{M}^0$.

In the general case, $\mathbf{t}^r(V)$ is the free algebra generated by $\mathbf{t}^{r-1}(V)$ and one indeterminate, so its multi-grading may be recursively defined similarly.

4.B.2 Dimensions of the truncated spaces

It is well known that if V is a d-dimensional space, then $V^{\otimes k}$ has dimension d^k. Hence we can write (Recalling Equation (4.10)):

$$\mathbf{ut}^1(V) = \bigoplus_{k \geq 0} \mathbf{ut}^1(V)_k, \quad \dim . \mathbf{ut}^1(V)_k = d^k.$$

In the case of $\mathbf{ut}^2(V)$ we may define $\mathbf{ut}^2(V)_k := \bigoplus_{\substack{\mathbf{k} \in \mathrm{Seq}^2, \\ \deg . \mathbf{k} = k}} \mathbf{ut}^2(V)_{\mathbf{k}}$ and write

$$\mathbf{ut}^2(V) = \bigoplus_{k \geq 0} \mathbf{ut}^2(V)_k, \quad \dim . \mathbf{ut}^2(V)_k = \frac{1}{2}(2d)^k.$$

To see why $\dim . \mathbf{ut}^2(V)_k = \frac{1}{2}(2d)^k$, note that since, as a vector space, $\mathbf{ut}^2(V)_{\mathbf{k}}$ is isomorphic to $\mathbf{ut}^1(V)_k$ and hence for any $\mathbf{k}$ with $\deg . \mathbf{k} = k$, it also has dimension d^k, so in order to determine the dimension of $\mathbf{ut}^2(V)_k$ it is enough to count $\#\{\mathbf{k} \in \mathrm{Seq}^2 : \deg . \mathbf{k} = k\} = 2^{k-1}$.

In the case of $\mathbf{t}^1(V)$ it is easily seen that $\dim . \mathbf{t}^1(V)_k = (d+1)^k$, hence we may write

$$\mathbf{t}^1(V) = \bigoplus_{k \geq 0} \mathbf{t}^1(V)_k, \quad \dim . \mathbf{t}^1(V)_k = (d+1)^k.$$

The case $\mathbf{t}^2(V)$ is slightly more complicated, but can be characterised by a simple linear recursion.

Proposition 4.B.2. *Let V be a d-dimensional vector space, then*

$$\mathbf{t}^2(V) = \bigoplus_{k\geq 0} \mathbf{t}^2(V)_k, \quad \dim.\mathbf{t}^2(V)_k := A_{d+1}(k),$$

where A_d satisfies the recursion

$$A_d(k) = (2d+1)A_d(k-1) - dA_d(k-2), \quad A_d(0) = 1,\ A_d(1) = d+1$$

Proof. Note that if $k \in \mathcal{M}^1$, then $\mathbf{t}^2(V)_k = \mathbf{t}^1(V)_k \simeq (V \oplus \mathbb{R})^{\otimes k}$ and if $k \in \mathcal{M}^1$, then $\mathbf{t}^2(V)_k \simeq \mathbb{R}$, putting this together we get for $\mathbf{k} = k_1 \cdots k_l \in \mathcal{M}(\mathcal{M}^1, \mathcal{M}^0)$

$$\dim \mathbf{t}^2(V)_{\mathbf{k}} = \prod_{k_i \in \mathcal{M}^1} \dim \mathbf{ut}^2(V)_{k_i} = (d+1)^{\sum_{k_i \in \mathcal{M}^1} k_i}.$$

Assume that $\mathbf{k} = k_1 \cdots k_n$ and $k_{i_1}, \ldots, k_{i_l} \in \mathcal{M}^1$, then since $\deg k_i = 1$ for any $i \neq i_1, \ldots, i_l$ it must be true that $\deg.\mathbf{k} = \sum_{k_i \in \mathcal{M}^1} k_i + n - l$. By setting $\mathbf{t}^2(V)_k := \bigoplus_{\substack{\mathbf{k} \in \mathcal{M}(\mathcal{M}^1, \mathcal{M}^0), \\ \deg.\mathbf{k} = k}} \mathbf{t}^2(V)_{\mathbf{k}}$ we see that

$$\dim \mathbf{t}^2(V)_k = \sum_{n=0}^{k} \sum_{l=0}^{n} \#\{\mathbf{k} = k_1 \cdots k_n \in \mathcal{M}(\mathcal{M}^1, \mathcal{M}^0) : \deg.\mathbf{k} = k, k_{i_1}, \ldots, k_{i_l} \in \mathcal{M}^1\}(d+1)^{l+k-n}.$$

We note that for n, l fixed we have

$$\begin{aligned}
&\#\{\mathbf{k} = k_1 \cdots k_n \in \mathcal{M}(\mathcal{M}^1, \mathcal{M}^0) : \deg.\mathbf{k} = k, k_{i_1}, \ldots, k_{i_l} \in \mathcal{M}^1\} \\
&= \binom{n}{l} \#\{\mathbf{k} = k_1 \cdots k_l \in \mathcal{M}^2 : \deg.\mathbf{k} = k + l - n\} \\
&= \binom{n}{l} \binom{k+l-n-1}{k-n}.
\end{aligned}$$

Hence

$$\dim \mathbf{t}^2(V)_k = 1 + \sum_{n=1}^{k} \sum_{l=1}^{n} \binom{n}{l} \binom{k+l-n-1}{k-n} (d+1)^{l+k-n}.$$

By summing over the diagonal this can be rewritten as

$$\begin{aligned}
\dim \mathbf{t}^2(V)_k &= 1 + \sum_{n=1}^{k} (d+1)^n \sum_{m=k-n+1}^{k} \binom{m}{m-k+n} \binom{n-1}{k-m} \\
&= 1 + \sum_{n=1}^{k} (d+1)^n (k+1-n)\, {}_2F_1(1-n, k+2-n, 2, -1),
\end{aligned}$$

where ${}_2F_1$ is the Gaussian Hypergeometric function. It follows that for $k \geq 2$

$$\begin{aligned}
&\dim \mathbf{t}^2(V)_k - (2d+3)\dim \mathbf{t}^2(V)_{k-1} + (d+1)\dim \mathbf{t}^2(V)_{k-2} \\
&= \sum_{n=2}^{k-1}(d+1)^n\Big[(k+1-n){}_2F_1(1-n,k+2-n,2,-1) - 2(k+1-n){}_2F_1(2-n,k+2-n,2,-1) \\
&\quad +(k-n){}_2F_1(2-n,k+1-n,2,-1) - (k-n){}_2F_1(1-n,k+1-n,2,-1)\Big] \\
&\quad +(d+1)^k\Big[{}_2F_1(1-k,2,2,-1) - 2{}_2F_1(2-k,2,2,-1)\Big] \\
&\quad +(d+1)\Big[k{}_2F_1(0,k+1,2,-1) - (k-1){}_2F_1(0,k,2,-1) - 1\Big].
\end{aligned}$$

Because of the two facts

$$ {}_2F_1(-k,2,2,-1) = 2^k, \quad {}_2F_1(0,k,2,-1) = 1,$$

all that remains is to show that for $a > 0$, $F(-a,b) := {}_2F_1(-a,b,2,-1)$ satisfies the recursion

$$(b+1)F(-a,b+2) - 2(b+1)F(1-a,b+2) + bF(1-a,b+1) - bF(-a,b+1) = 0.$$

To see this, note that by expanding ${}_2F_1(-a,b,2,-1)$ in its hypergeometric series we may write

$$ {}_2F_1(-a,b,2,-1) = \sum_{k=-\infty}^{\infty} f(a,b,k), \quad f(a,b,k) = \binom{a}{k}\frac{(b)_k}{(k+1)!}1_{\{0\leq k\leq a\}},$$

where $(b)_k$ is the rising Pochhammer symbol. It is straightforward to verify that f satisfies

$$\begin{aligned}
f(a+1,b,k) &= \frac{a+1}{a+1-k}f(a,b,k), \quad f(a,b+1,k) = \frac{b+k}{b}f(a,b,k), \\
f(a,b,k+1) &= \frac{(a-k)(b+k)}{(k+1)(k+2)}f(a,b,k).
\end{aligned}$$

By iterating these three relations one can show that

$$\begin{aligned}
&(b+1)f(a,b+2,k) - 2(b+1)f(a-1,b+2,k) + bf(a-1,b+1,k) - bf(a,b+1,k) \\
&= \frac{k(k+1)}{a}f(a,b+1,k) - \frac{(k+1)(k+2)}{a}f(a,b+1,k+1),
\end{aligned}$$

and the claimed recursion follows by summing over k since

$$\begin{aligned}
&(b+1)F(-a,b+2) - 2(b+1)F(1-a,b+2) + bF(1-a,b+1) - bF(-a,b+1) \\
&= \sum_{k=-\infty}^{\infty}(b+1)f(a,b+2,k) - 2(b+1)f(a-1,b+2,k) + bf(a-1,b+1,k) - bf(a,b+1,k) \\
&= \sum_{k=-\infty}^{\infty}\frac{k(k+1)}{a}f(a,b+1,k) - \sum_{k=-\infty}^{\infty}\frac{(k+1)(k+2)}{a}f(a,b+1,k+1) = 0.
\end{aligned}$$

□

Remark 4.B.2. The sequences $A_0(k), A_1(k)$ are listed as $A001519$ and $A052984$ respectively on OEIS.

4.B.3 Higher rank tensor algebras on Banach spaces

Definition 35. We make $\mathbf{t}^r(V)$ into a normed space with the norm defined inductively as

$$\|t\|_r = \sum_{m\geq 0} \|\pi_m t\|_{\mathbf{t}^{r-1}(V)^{\otimes m}}$$

where $\pi_m : \mathbf{t}^r(V) \to \bigoplus_{\deg.\mathbf{k}=m} \mathbf{t}^{r-1}(V)_{\mathbf{k}}$ denotes the projection map onto components of degree k. Define $\mathbf{T}^r(V)$ to be the completion of $\mathbf{t}^r(V)$ under the norm $\|\cdot\|_r$.

Remark 4.B.3. Since the embedding $\mathbf{t}^r(V) \hookrightarrow \mathbf{t}^{r+1}(V)$ is an isometric isomorphism onto its image, the same is true for the embedding $\mathbf{T}^r(V) \hookrightarrow \mathbf{T}^{r+1}(V)$.

Remark 4.B.4. Note that S_r indeed takes values in $\mathbf{T}^r(E)$, since by the above Remark 4.B.3 it is enough to show that S_1 takes values in $\mathbf{T}^1(E)$ which follows from multiplication and addition being continuous and the exponential series being absolutely convergent.

By unravelling Definition 35 we may write for $t \in \mathbf{t}^r(V)$

$$\|t\|_r = \sum_{\mathbf{k}\in\mathcal{M}^r} \|\pi_{\mathbf{k}} t\|$$

where $\pi_{\mathbf{k}} : \mathbf{t}^r(V) \to \mathbf{t}^r(V)_{\mathbf{k}}$ is projection onto $\mathbf{t}^r(V)_{\mathbf{k}}$ which is topologically isomorphic to a tensor copy of V, hence it has a well defined norm by the assumption that V has an admissible norm. Finally, we note that if V is a Hilbert space, then $\mathbf{T}^r(V)$ also possesses a Hilbert space structure.

Definition 36. For a Hilbert space V we equip $\mathbf{ut}^r(V)$ with the recursively defined inner product

$$\langle t, s\rangle_r = \sum_{m\geq 0} \langle \pi_m t, \pi_m s\rangle_{\mathbf{ut}^{r-1}(V)^{\otimes m}}$$

and we denote by $\tilde{\mathcal{H}}^r(V)$ and $\mathcal{H}^r(V)$ the respective completions of $\mathbf{ut}^r(V)$ and $\mathbf{t}^r(V)$ with this inner product.

4.B.4 Tensor normalization estimates

Recall that the scaling of an element $v \in V$ by $\lambda \in \mathbb{R}$, $\lambda \mapsto \lambda v$, extends naturally to a dilation map on $\prod_{m\geq 0} V^{\otimes m}$:

$$\delta_\lambda : \mathbf{t} \mapsto (\mathbf{t}^0, \lambda \mathbf{t}^1, \lambda^2 \mathbf{t}^2, \ldots).$$

Definition 37. A tensor normalization is a continuous injective map of the form

$$\Lambda : \mathbf{T}^1(V) \to \{\mathbf{t} \in \mathbf{T}^r(V) : \|\mathbf{t}\| \leq 1\}, \quad \mathbf{t} \mapsto \delta_{\lambda(\mathbf{t})}\mathbf{t},$$

where $\lambda : \mathbf{T}^r(V) \to (0, \infty)$ is a function.

It is possible to show that there always exists a tensor normalization, see [46, Proposition A.2 and Corollary A.3].

Theorem 4.B.1. *For any Banach space V and any admissible norm on $(V^{\otimes m})_{m\geq 1}$, there exists a tensor normalization map Λ.*

For any (discrete time) path $x \in (I \to V)$, $\mathrm{S}(x)$ takes values in $\mathbf{T}^1(V)$, see [131, Theorem 3.12]. In particular, for a given tensor normalization Λ, $\Lambda \circ \mathrm{S}(x)$ takes values in the unit ball of the Banach space $\mathbf{T}^1(V)$, and therefore for any $\mu \in \mathrm{Meas}(I \to V)$, the Bochner integral $\overline{\Lambda \circ \mathrm{S}}(x) := \int_{x \in V^I} \Lambda \circ \mathrm{S}(x) \mu(dx)$ is well–defined. Then we may iteratively define $\Lambda\,\mathrm{S}^r := \Lambda \circ \mathrm{S} \circ \Lambda\,\mathrm{S}^{r-1}$

Definition 38. We call $\mathrm{S}^n_r := \Lambda\,\mathrm{S}^r$ the robust (or, normalized) signature map of rank r and $\bar{\mathrm{S}}^n_r := \mathbb{E}\,\mathrm{S}^n_r$ is called the robust expected signature map of rank r.

The proof of the next proposition can be found in [46, Corollary 5.7].

Proposition 4.B.3. *Let V be a separable Banach space. Then $\bar{\mathrm{S}}^n_1 : \mathrm{Meas}(I \to V) \to \mathbf{T}^1(V)$ is injective.*

For our concrete purpose in Sect. 4.4.3, we introduce the following robust signature, which is slightly different from the one we defined above as it is not of the form $\Lambda \circ \mathrm{S}$.

Proposition 4.B.4. *Let V be a separable Banach space. Let $\Phi : (I \to V) \to \mathbf{T}^1(V)$ be the map such that for $x \in (I \to V)$,*

$$\Phi(x) = \delta_{\exp(-\|\mathrm{S}(x)\| - \|x\|_\infty)}\,\mathrm{S}(x).$$

Then Φ is bounded continuous and injective.

Proof. Let $\mathbb{1}$ denote the neutral element $(1, 0, \ldots)$ in $\mathbf{T}^1(V)$ with respect to the tensor product.
The boundedness of Φ is clear, because for any $a \in [0, 1]$ it holds that $\|\delta_a \,\mathrm{S}(x) - \mathbb{1}\| \leq a\|\,\mathrm{S}(x)\|$, inserting $a = \exp(-\|\,\mathrm{S}(x)\| - \|x\|_\infty)$ we indeed get a uniform bound for Φ. To show the continuity of Φ, note that for x^n converges to x in $I \to V$, we have

$$\begin{aligned}\|\Phi(x^n) - \Phi(x)\| \leq &\|\delta_{\exp(-\|\,\mathrm{S}(x^n)\| - \|x^n\|_\infty)}\,\mathrm{S}(x^n) - \delta_{\exp(-\|\,\mathrm{S}(x^n)\| - \|x^n\|_\infty)}\,\mathrm{S}(x)\| \\ &+ \|\delta_{\exp(-\|\,\mathrm{S}(x^n)\| - \|x^n\|_\infty)}\,\mathrm{S}(x) - \delta_{\exp(-\|\,\mathrm{S}(x)\| - \|x\|_\infty)}\,\mathrm{S}(x)\|.\end{aligned}$$

The first term on the right hand side converges to 0 as it is bounded by $\|\,\mathrm{S}(x^n) - \mathrm{S}(x)\|$ which vanishes as $n \to \infty$ by the continuity of S; the second term on the right hand side also tends to 0 by dominated convergence as $\mathrm{S}(x)$ has a factorial decay in its tail (cf. [131, Theorem 3.1.2]). For the injectivity of Φ, let us assume that $\Phi(x) = \Phi(y)$ for $x, y \in I \to V$. Using the relation that $\delta_a \delta_b \,\mathrm{S}(x) = \delta_{ab}\,\mathrm{S}(x)$ for all $a, b \in \mathbb{R}$ it implies that

$$\delta_c \,\mathrm{S}(x) = \mathrm{S}(y),$$

where $c = \exp(\|\,\mathrm{S}(y)\| + \|y\|_\infty - \|\,\mathrm{S}(x)\| - \|x\|_\infty)$. Then by [78, Exercise 7.55] we have $\delta_c(\mathrm{S}(x)) = \mathrm{S}(cx) = \mathrm{S}(y)$. However, keeping in mind that we included time component into the definition of S, it holds that the projection of $\mathrm{S}(cx)$ to the first level $\mathbb{R} \oplus V$ is equal to (cT, cx_T) while the counterpart for $\mathrm{S}(y)$ is (T, y_T). Therefore we must have $cT = T$, i.e., $c = 1$. Consequently it follows that $\mathrm{S}(x) = \mathrm{S}(y)$ and also $x = y$ by the injecvitity of S (see Theorem 4.3.1). □

Remark 4.B.5. Note that the injectivity of Φ depends crucially on the fact that we include time component into the definition of signature map, and it may not be true if one uses signature without time extension. In the latter case one has to apply the tensor normalization introduced in [46]. Also note that we include $\|x\|_\infty$ into the dilation for a special technical reason, see discussion in the next section.

4.C Feature Maps, MMDs and Weak Convergence

In this Section we provide the necessary background for the robust signature map S_r^n that we use to deal with non-compactness, see Section 4.3.5. Central to our argument is to exploit a duality between functions and measures via a "universal feature map". In the non-compact case this duality can be subtle to handle, see [169] for an overview.

4.C.1 Universality and Characteristicness

Definition 39. Let $\mathcal{X}$ be a topological space and E be a topological vector space. We call any map $\Phi; \mathcal{X} \to E$ a feature map. Moreover, for a given topological vector space $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$, we say that a feature map Φ is

1. universal to $\mathcal{F}$, if the map

$$\iota : E' \to \mathbb{R}^{\mathcal{X}}, \quad \ell \mapsto \langle \ell, \Phi(\cdot) \rangle$$

   has a dense image in $\mathcal{F}$, where E' denotes the topological dual of E.

2. characteristic to a subset $\mathcal{P} \subset \mathcal{F}'$ if the map

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

   is injective, where $(E')^*$ denotes the algebraic dual of E'.

The following duality is a direct consequence of the Hahn–Banach Theorem, see e.g. [46, Theorem 2.3].

Theorem 4.C.1. *If $\mathcal{F}$ is a locally convex space, then a feature map Φ is universal to $\mathcal{F}$ if and only if Φ is characteristic to $\mathcal{F}'$.*

4.C.2 Robust Signature Features and their Topology

Put in our context, the feature map is the (robust) signature map S_r^n.

Theorem 4.C.2. *Define $\Delta(v) := 1 \otimes v + v \otimes 1$ for $v \in V$. Then $(\mathbf{T}^1(V), \otimes, \Delta)$ is a Hopf algebra and the co-domain of both the signature map S and the robust signature map S_1^n is the set of group-like elements*

$$G := \{g \in \mathbf{T}^1(V) : \Delta g = g \otimes g\} \subset \mathbf{T}^1(V),$$

that is

$$\mathrm{S}, \mathrm{S}_1^n \ : \ (I \to E) \to G \subset \mathbf{T}^1(V).$$

Moreover, both these maps are continuous and injective.

Proof. This is classical for S and follows for S_1^n by integration by parts, see [46, Sect. 5.1]. □

The following proposition is crucial for this present paper.

Proposition 4.C.1. *Assume that $\mathcal{X}$ is metrizable. Then for any continuous injective mapping $\varphi : \mathcal{X} \to V$, where V is a Banach space, the map*

$$\Phi := \mathrm{S}_1^n \circ \varphi : \mathcal{X}^I \to \mathbf{T}^1(V)$$

is universal to $C_b(\mathcal{X}^I, \mathbb{R})$ and characteristic to $C_b(\mathcal{X}^I, \mathbb{R})'$. In particular, two finite regular Borel measures μ and ν on $\mathcal{X}^I$ are equal if and only if $\int \Phi d\mu = \int \Phi d\nu$.

Proof. Let $x \in (I \to \mathcal{X})$ be a (discrete time) path taking values in $\mathcal{X}$. Then $\varphi \circ x$ is a (discrete time) path taking values in V. Thanks to Theorem 4.C.2 the map $\Phi = \mathrm{S}_1^n \circ \varphi$ is continuous and injective, and takes values in $G \subset \mathbf{T}^1(V)$. Define $L := \bigoplus_{m \geq 0} (V^{\otimes m})^\star$, which we identify with a dense subspace of $\mathbf{T}^1(V)^\star$ via $\ell(\mathbf{t}) = \sum_{m \geq 0} \langle \ell^m, \mathbf{t}^m \rangle$, and define $\tilde{\mathcal{F}} := \{\ell \circ \Phi : \ell \in L\}$. Clearly, $\tilde{\mathcal{F}} \subset \mathcal{F} = C_b(\mathcal{X}^I, \mathbb{R})'$, and the injectivity of Φ implies that $\tilde{\mathcal{F}}$ separates the points in $\mathcal{X}^I$. Furthermore, the algebraic condition on G implies that $\tilde{\mathcal{F}}$ is closed under multiplication (when $V = \mathbb{R}^d$, this is equivalent to the shuffle product equation in [129, (2.6)]). This implies that $\tilde{\mathcal{F}}$ satisfies all conditions in [46, Theorem 2.6, (2)] and is therefore dense in $C_b(\mathcal{X}^I, \mathbb{R})$ with respect to the strict topology by [46, Theorem 2.6]. This means that $\tilde{\mathcal{F}}$ is universal to $C_b(\mathcal{X}^I, \mathbb{R})$ and characteristic to $C_b(\mathcal{X}^I, \mathbb{R})'$ by [46, Theorem 2.3]. The last assertion then follows immediately. □

4.C.3 Kernelized Maximum Mean Discrepancies

Following [46] we now use S_1^n to define a kernel and show that the associated Maximum Mean Discrepancy (MMD) metrizes weak convergence when the state space $V = \mathbb{R}^d$.

Proposition 4.C.2. *Let $V = \mathbb{R}^d$ Define*

$$\mathbf{k} : (I \to V) \times (I \to V) \to \mathbb{R}, \quad \mathbf{k}(x, y) := \langle \mathrm{S}_1^n(x), \mathrm{S}_1^n(y) \rangle - 1.$$

Then

1. *$\mathbf{k}$ is a continuous, bounded, positive definite function.*
2. *$\mathbf{k}$ is characteristic to $C_b(I \to V)'$.*
3. *the reproducing kernel Hilber space (RKHS) generated by $\mathbf{k}$, $\mathcal{H}_{\mathbf{k}}$, is a subset the space of all continuous functions on $I \to V$ vanishing at infinity,*

$$\mathcal{H}_{\mathbf{k}} \subset C_0(I \to V).$$

Proof. The first statement follows immediately from Proposition 4.B.4 as S_r^n is a bounded continuous mapping with values in $\mathcal{H}^1(V)$. To see characteristicness, note that $\mathbf{k}(x,y) = \langle \mathrm{S}_1^n(x) - \mathbb{1}, \mathrm{S}_1^n(y) - \mathbb{1}\rangle$ for $\mathbb{1} = (1,0,\dots) \in \mathcal{H}^1(V)$ the unit element. By [46, Proposition 7.3] it then remains to show that $\tilde{\mathrm{S}}_1^n(x) := \mathrm{S}_1^n(x) - \mathbb{1}$ is characteristic to $C_b(I \to V)'$. However, this property is inherited from the corresponding property of S_1^n, Proposition 4.C.1. To be precise, the construction of S_1^n ensures that $\mathrm{S}_1^n(x)$ is group–like. Consequently $\{\langle \ell, \mathrm{S}_1^n(x)\rangle : \ell \in \mathbf{t}^1(V)\}$ forms an algebra, which in turn implies that $\{\langle \ell, \tilde{\mathrm{S}}_1^n(x)\rangle : \ell \in \mathbf{t}^1(V)\}$ is also an algebra as $\tilde{\mathrm{S}}_1^n(x)$ coincides with $\tilde{\mathrm{S}}_1^n(x)$ on $\Pi_{m=1}^\infty (\mathbb{R}\oplus V)^{\otimes m}$ and the projection of $\tilde{\mathrm{S}}_1^n(x)$ to $(\mathbb{R}\oplus V)^{\otimes 0} \sim \mathbb{R}$ equals 0. Furthermore, the boundedness and continuity of S_1^n ensures that $\{\langle \ell, \tilde{\mathrm{S}}_1^n(x)\rangle : \ell \in \mathbf{t}^1(V)\} \subset C_b(I \to V)$; the injectivity guarantees that $\{\langle \ell, \tilde{\mathrm{S}}_1^n(x)\rangle : \ell \in \mathbf{t}^1(V)\}$ separates points; finally, since each $\tilde{\mathrm{S}}_1^n(x)$ contains time component $\exp(-\|\mathrm{S}(x)\| - \|x\|_\infty)T \neq 0$ as we are using time extended signature, the set $\{\langle \ell, \tilde{\mathrm{S}}_1^n(\cdot)\rangle : \ell \in \mathbf{t}^1(V)\}$ still contains constant functions. Hence, we can use a Stone–Weierstrass type argument as in Proposition 4.C.1, see also [46, Theorem 2.6], to deduce that $\tilde{\mathrm{S}}_1^n(\cdot)$ is characteristic to $C_b(I \to V)'$.

Finally, note that for $x \in I \to V$, one has $\lim_{\|y\|_\infty \to \infty} |\mathbf{k}(x,y)| = 0$, because

$$\begin{aligned}
|\mathbf{k}(x,y)| = |\langle \tilde{\mathrm{S}}_1^n(x), \tilde{\mathrm{S}}_1^n(y)\rangle| &\leq \|\tilde{\mathrm{S}}_1^n(x)\|\|\tilde{\mathrm{S}}_1^n(y)\| \\
&= \|\tilde{\mathrm{S}}_1^n(x)\| \left\| \delta_{\exp(-\|\mathrm{S}(y)\| - \|y\|_\infty)}\, \mathrm{S}(y) - \mathbb{1} \right\| \\
&\leq \|\tilde{\mathrm{S}}_1^n(x)\| \exp(-\|\mathrm{S}(y)\| - \|y\|_\infty) \|\mathrm{S}(y)\| \\
&\to 0
\end{aligned}$$

as $\|y\|_\infty \to \infty$. Hence, in view of [169, Lemma 4.1] one can conclude that $\mathcal{H}_{\mathbf{k}} \subset C_0(I \to V)$. □

We now conclude by [169, Lemma 2.1]

Corollary 4.C.3. *Let $V = \mathbb{R}^d$. Then*

$$d_{\mathbf{k}}(\mu,\nu) = \left\| \int \mathrm{S}_r^n(x)\mu(dx) - \int \mathrm{S}_r^n(x)\nu(dx) \right\| = \left\| \bar{\mathrm{S}}_1^n(\mu) - \bar{\mathrm{S}}_1^n(\nu) \right\|$$

characterizes weak convergence.

Chapter 5

Proper Scoring Rules, Gradients, Divergences, and Entropies for Paths and Time Series

This chapter is based on [29] which was written with Harald Oberhauser.

We study scoring rules to assess forecasts of trajectories, given either in discrete or continuous time. Our approach leverages the statistical framework of proper scoring rules with classical mathematical results to derive a principled approach to decision making with such forecasts. In particular, we introduce notions of gradients, entropy, and divergence that are tailor-made to respect the underlying non-Euclidean structure. This results in computable quantities and we demonstrate their efficiency in classical inference tasks such as two-sample and independence testing.

5.1 Introduction

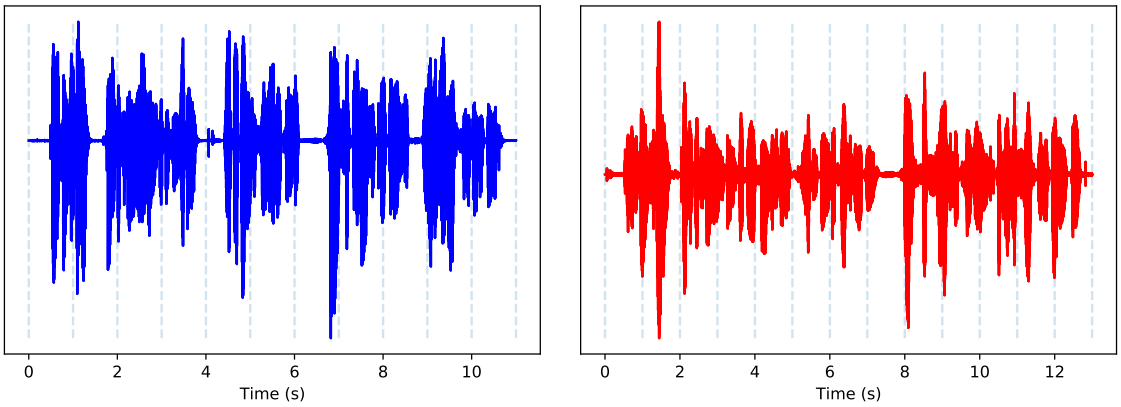
Scoring rules provide a principled approach to form and evaluate probabilistic predictions. The earliest applications go at least back to the evaluation of weather forecasts [32], but have since then developed into a rich theoretical framework that plays a central part in modern statistical inference. We refer to [84] and [57] for a general background. The theoretical underpinnings of scoring rules are well-developed, but nearly all of the literature focuses on prediction of vector-valued or scalar-valued quantities. The aim of this article is to develop a scoring rule framework for an important class of non-Euclidean data, namely sequential data – both in discrete and continuous time. That is a forecaster gives us a probability measure on the space of paths (continuous time) or sequences (discrete time) and we want to develop a scoring rule to assess such forecasts about the future trajectory of a quantity.

The Drawbacks of Naïve Vectorization. A common approach is to flatten sequential data into a long vector and then apply a (scoring rule) pipeline for vector-valued data[1]. However, this approach has several drawbacks:

1. **Irregular Sampling.** Firstly, there's trouble whenever the time series are irregularly sampled or are of different length since this embeds the different time series in Euclidean spaces of different dimension. Typically this is addressed by ad-hoc approaches such as adding synthetic data, interpolation, or dropping of data points.

2. **Parametrization (in)-variance.** Secondly, often the relevant information is independent of the time-parametrization ("*time-warping invariance*"), at least to a large degree; for instance, the meaning of a spoken word or an object being filmed are both independent of how fast or slow the audio signal or video is presented. As an example of this, in Figure 5.1 we show two different individuals reading the same sentences. Their intonation and natural cadence are very different resulting in clearly distinct audio samples. Semantically, these two files are the same, they carry the same meaning. However, by identifying them as vectors where each time sample is vector coordinate, these two sentence would be very far apart in Euclidean metric. In fact, they are not even in the same space since one person takes 13 seconds to complete the sentence and the other takes 11 seconds. This is a common motivation for algorithms like Dynamic

[1] Given $x(t_1), \dots, x(t_L) \in \mathbb{R}^d$ it is flattened to the vector in $\mathbb{R}^{dL}$ which has as first d coordinates just the coordinates of $x(t_1) \in \mathbb{R}^d$, the following d coordinates are those of $x(t_2) \in \mathbb{R}^d$, etc.



(a) Speaker from Glasgow, Scotland. (b) Speaker from Yole, India.

Figure 5.1: Two speakers saying *"Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob."*. These samples are from the speech accent archive [201].

time warping (DTW), as it lines up the two signals in a way that is invariant under the rhythm of the individuals speaking.

3. **Continuous-Time Models.** Many models are naturally formulated in continuous time rather than discrete time; for example, stochastic differential equations form a popular class of models in many applications and it is unclear how to evaluate such continuous time models in a scoring rule framework besides above naive vectorization on a chosen time grid. Ideally, we want a scoring rule framework that is general enough to handle both the discrete and continuous-time forecasts. For a demonstration of how the scoring rule depends on the granularity of the time-discretization, see Figure 5.7.

A Non-Euclidean Data Domain. For simplicity we now focus on continuous time, since discrete time can be treated as a special case[2]. That is by a forecast given observations up to time t we mean a probability measure on the space $\{x|x : [t, T] \to \mathbb{R}, x \text{ continuous}, T > 0\}$. The latter is not a linear space since there is no natural addition for paths on different domains, $[t, T_1]$ and $[t, T_2]$. Even worse, as Figure 5.1 shows often one would like to forget the time-parametrization, that is identify $t \mapsto x(t)$ and $t \mapsto x(\varphi(t))$ where φ is any increasing function (a "time change"). Formally, one

[2]Given discrete-time observations $x(t_1), \dots, x(t_n)$, we identify them as the continuous time path constructed by piecewise linear interpolation, formally $t \mapsto x(t_i) + \frac{t-t_i}{t_{i+1}-t_i}(x(t) - x(t_i))$ for $t \in [t_i, t_{i+1})$. See also the end of Section 5.3 for more details.

works with a space of equivalence classes of paths rather than paths. We refer to such an equivalence class of paths as a *track*, as it is defined uniquely by the track it carves out in the space where it evolves. The space of paths and the space of tracks are not linear but exhibit a rich structure: any two paths (resp. tracks) can be concatenated into one path and any path (resp. track) can be run backwards. Key to our approach is that classic tools from pure mathematics [41] faithfully capture this Non-Euclidean structure by turning these operations of concatenation and time-reversal into algebraic operations. This allows us to define a scoring rule that respects this Non-Euclidean structure. Besides probability measures on the space of paths resp. tracks one is also interested in functions on these spaces (e.g. the running average $\frac{1}{T-t}\int_t^T x(s)ds$ of a path). The usual definition of differentiability does not apply due to the above lack of linear structure. We address this by revisiting classic work of Pansu [143] which we use to define *gradients of functions of (unparametrized) paths*. This gradient structure further allows us to compute quantities associated with our scoring rule framework via gradient descent; even in classical scoring rule framework for Euclidean data, many quantities are typically not given in closed form, but can be found by first order methods and the same situation arises here.

Outline. Section 5.2 recalls the basic definitions and general properties of scoring rules. Section 5.3 contains the theoretical background; it formalizes the structure of the spaces of paths, resp. tracks, and introduces the signature feature map. Section 5.4 uses this setting to derive natural scoring rules; we do this by leveraging that the signature turns operations on paths and tracks into algebraic operations in feature space. From general results, this leads to definitions of entropy, divergence, and mutual information that – unlike the (naïve) vectorization approach outlined above – are compatible with the structure of (probability measures) on spaces of paths and tracks. Section 5.5 develops the concept of differentiability; since our data domain is not Euclidean, classical differentiability does not apply but we show that so-called Pansu differentiability leads to a natural notion of gradients. This allows us to develop gradient descent in this context which allows us to compute many of the quantities that arise in our scoring rule framework. Finally, Section 5.6 provides experiments that show that despite this approach being motivated by pure maths, the resulting quantities lead to efficiently computable quantities that show favourable performance and are simple to use in a modular framework like any scoring rule.

5.1.1 Related Work

One of the earliest empirical insights for time series data was that time-parametrization ("time warping") invariance is of key importance [159, 158]. Arguably the most popular way to address this invariance is via the classical dynamic time warping distance (DTW) and its many variations that introduce a distance between time series by searching over time changes. For example, [53, 22] introduce a regularised version of DTW, so-called soft DTW (s-DTW) which addresses the fact that the DTW distance is not differentiable, making it viable for use in deep learning pipelines. In the process of doing so it loses the invariance that DTW enjoys and introduces a trade-off of smoothness versus invariance. Another drawback is that that DTW computations scale quadratically in the number of sequence entries and this makes it too expensive for many sources of high-frequency data. In contrast, our distance can be computed in linear time for the price of higher complexity in the state space dimension. A more general point is that DTW and its variations do not aim to provide the full forecasting framework of scoring rules (divergence between measures, entropy, mutual information of time series); further, although DTW approaches successfully deal with time-parametrization, they ignore other structural properties such as concatenation and reversal of times series.

Another directly related approach is kernel learning. Any kernel $k : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ with reproducing kernel Hilbert space H_k induces a scoring rule by setting $s(x, \mu) := \|\delta_x - \mu\|^2_{H_k}$ where $\|\mu\|^2_{H_k} = \int\int k(x, y)\mu(dx)\mu(dy)$, see [84, Section 5] and several kernels for sequences have been developed in the literature. Most relevant to our approach is the "signature kernel" introduced in [113]. However, for any scoring rule given by a kernel, the (generalized) divergence becomes simply the maximum mean discrepancy and the entropy simply the variance in the RKHS H_k. While kernels give rise to a powerful class of scoring rules, the success and popularity of non-kernel based scoring rules on $\mathcal{X} = \mathbb{R}^n$ is motivation enough to look for other interesting, non-linear scoring rules.

The technical key to our approach comes from mathematics where iterated integrals, so-called signatures, and non-commutative algebras are use to represent paths. This goes at least back to seminal work of Chen [41] in algebraic topology and subsequent applications in control theory [71, 34] and more recently rough path theory [129]. These results have been influential in stochastic analysis [131, 76] and only more recently have been started to be explored in a statistical and machine learning context. We mention pars-pro-toto [144, 46, 28] for inference about laws of stochastic processes; [113, 160, 67] for kernel learning; [190, 62, 121] for Bayesian approaches; [140, 19, 110]

for generative modelling; [83, 119, 45] for applications in topology; [59, 60, 189] for algebraic perspectives.

Finally, we mention that the two topics that are central to us – invariances and non-Euclidean structure – have been considered in different contexts in scoring rule frameworks. For example, [70] studies equi- and in-variances for scoring rules for Euclidean data; non-vector valued data such as sets, contours, intervals, and quantiles have received attention [25, 134, 69].

5.2 Proper Scoring Rules, Entropies, and Divergences

We briefly recall general background on scoring rules following closely the notation in [57], see also [84, 56, 95, 91]. Let $\mathcal{X}$ be a measurable space (*the outcome space*), $\mathcal{A}$ be a set (*the action space*), and $\mathcal{L} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ be a function (*the loss function*). Further, let X be a $\mathcal{X}$-valued random variable. We consider the following game between a Decision-maker and Nature: the task of the decision maker is to choose an action $a \in \mathcal{A}$, after which Nature reveals the outcome $x \in \mathcal{X}$ that is given by sampling X. The decision maker then suffers the loss $\mathcal{L}(x, a)$.

Given a set $\mathcal{P}$ of probability measures on $\mathcal{X}$, a principled probabilistic (Bayesian) approach to this decision problem is to proceed as follows:

1. Associate with every $\mu \in \mathcal{P}$ its *Bayes act* $a_\mu \in \mathcal{A}$ defined by

$$a_\mu := \arg\min_{a \in \mathcal{A}} \mathbb{E}_{X \sim \mu}[\mathcal{L}(X, a)]$$

   (assuming that a minimum exists; if it is not unique, choose a_μ arbitrary among the minimizers).

2. Use the Bayes act a_μ to define the *scoring rule* $s : \mathcal{X} \times \mathcal{P} \to \mathbb{R}$ on $\mathcal{X}$ given by

$$s(x, \mu) := \mathcal{L}(x, a_\mu), \tag{5.1}$$

3. Use the scoring rule s to define the (generalised) *entropy* H, the (generalised) *divergence* $d : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$, and the (generalised) *mutual information* $I : \mathcal{P} \times \mathcal{Q} \to \mathbb{R}$ as

$$\begin{aligned} H : \mathcal{P} \to \mathbb{R}, &\quad \mu \mapsto \mathbb{E}_{X \sim \mu}[s(X, \mu)], \\ d : \mathcal{P} \times \mathcal{P} \to \mathbb{R}, &\quad (\nu, \mu) \mapsto \mathbb{E}_{X \sim \nu}[S(X, \mu)] - H(\nu), \\ I : \mathcal{P} \times \mathcal{Q} \to \mathbb{R}, &\quad (\mu, \nu) \mapsto H(\mu) - \mathbb{E}_{U \sim \nu}[H(\mu|U)]. \end{aligned}$$

   where $\mathcal{Q}$ denotes a space of probability measures (not necessarily on $\mathcal{X}$) and $\mu|U$ denotes the law of μ conditioned on the random variable U.

Table 5.1: Table of symbols used.

Symbol	Meaning	Page
	Scoring rules	
$\mathcal{X}$	a measurable space	169
$\mathcal{A}$	any set representing an action space	169
$\mathcal{L}$	a loss function $\mathcal{A} \times \mathcal{X} \to \mathbb{R}$	169
$\mathcal{P}, \mathcal{Q}$	a set of probability measures	169
a_μ	the Bayes' act in $\mathcal{A}$ associated to the measure μ	169
$s(\cdot, \cdot)$	a scoring rule $\mathcal{X} \times \mathcal{P} \to \mathbb{R}$	169
$H(\cdot)$	the generalised entropy $\mathcal{P} \to \mathbb{R}$ associated to a scoring rule	169
$d(\cdot, \cdot)$	the generalised divergence $\mathcal{P} \times \mathcal{P} \to \mathbb{R}$ associated to a scoring rule	169
$I(\cdot, \cdot)$	the generalised mutual information $\mathcal{P} \times \mathcal{Q} \to \mathbb{R}$	169
	Tensors and Paths	
$\otimes$	denotes the tensor product between vector spaces and their elements	171
$\mathcal{H}$	the tensor algebra equipped with its Hilbert space norm.	172
Paths	the space of paths.	174
Tracks	the space of unparametrised paths.	173
Φ	the feature map Tracks $\to \mathcal{H}$.	173
$\exp(\cdot)$	the tensor exponential $\exp(v) := (1, v, \frac{v^{\otimes 2}}{2!}, \ldots) \in \mathcal{H}$.	177
α	the antipode $\mathcal{H} \to \mathcal{H}$.	180
G	the set of grouplike elements in $\mathcal{H}$.	177
$\mathcal{H}^\star$	the elements in $\mathcal{H}$ starting with a 1.	179
L	a convex function $\mathcal{H} \to \mathbb{R}$ with a unique minimum at the unit	179
$\mathcal{L}^{\text{left}}$	the map Tracks $\times \mathcal{H}^\star \to [0, \infty)$, $\quad (\mathbf{x}, m) \mapsto L(m^{-1}\Phi(\mathbf{x}))$	179
	Operations on Measures and Paths	
$\mathbf{x}, \mathbf{y}$	Two paths	174
μ, ν	probability measures in $\mathcal{P}$	169, 171
$\mathbf{x} \star \mathbf{y}$	The concatenated path of $\mathbf{x}$ and $\mathbf{y}$	174
$\overleftarrow{\mathbf{x}}$	The reverse path of $\mathbf{x}$	174
$\mu \star \nu$	The law of the concatenated paths of $\mathbf{x} \star \mathbf{y}$ where $\mathbf{x} \sim \mu, \mathbf{y} \sim \nu$	180
$\overleftarrow{\mu}$	The law of $\overleftarrow{\mathbf{x}}$ where $\mathbf{x} \sim \mu$	180
$\perp\!\!\!\perp$	denotes independence between two measures	171
$\mu \vert \nu$	denoted the measure μ conditioned on ν	169

The above definitions and nomenclature is justified as follows: firstly, it is an instructive exercise to check that for standard choices of state space $\mathcal{X}$, action space $\mathcal{A}$, and loss function $\mathcal{L}$, the above reduce to classical definitions of entropy, divergence and mutual information (e.g. if $\mathcal{A}$ is the set of densities on $\mathbb{R}^n$ then the *log score* $\mathcal{L}(x, a) := -\log a(x)$ from [86] yields the usual Shannon entropy, Kullback-Leibler divergence); see [57] for more examples. Secondly, Theorem 5.2.1 below shows that characteristic properties hold in the full generality of the above setup:

Theorem 5.2.1 ([57])**.** *Let $\mathcal{X}$, $\mathcal{A}$, and $\mathcal{L}$ be as above. Further, denote with H, I, and d the associated (generalized) entropy, mutual information, and divergence. Then*

1. *the scoring rule* (5.1) *is* proper*, that is*

$$\mu \mapsto \mathbb{E}_{X\sim\nu}[s(X, \mu)]$$

   is minimized at $\mu = \nu$.

2. *$\mu \mapsto H(\mu)$ is concave,*

3. *$(\nu, \mu) \mapsto \mathbb{E}_{X\sim\nu}[s(X, \mu)]$ is affine in ν for every μ,*

4. *$d(\mu, \nu) \geq 0$ with equality for $\mu = \nu$,*

5. *$I(\mu, \nu) \geq 0$ with equality for $\mu \perp\!\!\!\perp \nu$.*

Different applications areas demand different scoring rules. Classical choices for the Euclidean case $\mathcal{X} = \mathbb{R}^n$ are besides the already mentioned log score, the Brier score, the Tsallis score, the Bregman score, the Hyvärinen score, etc.; see [57] for details. The aim of the remainder of this article is to study the case when $\mathcal{X}$ is a space of paths or a space of equivalence classes of paths (under reparametrisation).

A Toy Example: From Feature Maps to Bayes Actions. To motivate our scoring rule for paths let us first revisit the vector-valued case, $\mathcal{X} = \mathbb{R}^n$. To arrive at a proper scoring rule, the space of Bayes actions $\mathcal{A}$ should be large enough to characterize any (sufficiently nice) probability measures on $\mathcal{X} = \mathbb{R}^n$. A classic way to characterize a probability measure, is to consider the sequence of moments,

$$(1, \mathbb{E}[X], \mathbb{E}[X^{\otimes 2}], \mathbb{E}[X^{\otimes 3}], \ldots) \tag{5.2}$$

that is, $\mathbb{E}[X]$ is the mean vector, $\mathbb{E}[X^{\otimes 2}]$ is the covariance matrix, etc. (We tacitly assume that the sequence of moments is well-defined and decays quickly enough so

that it characterizes the measure, see Remark 5.2.1). The sequence (5.2) is an element of the set

$$\mathcal{H} := \prod_{m \geq 0} (\mathbb{R}^n)^{\otimes m} \tag{5.3}$$

of sequences of tensor of increasing degree m. In fact, this set $\mathcal{H}$ forms a vector space by element-wise addition of tensors of the same degree. We define the "feature map"

$$\varphi : \mathbb{R}^n \to \mathcal{H}, \quad x \mapsto (1, x, x^{\otimes 2}, \ldots). \tag{5.4}$$

With the above notation, the moment sequence (5.2) is simply the mean of $\varphi(X)$,

$$\mathbb{E}[\varphi(X)] = (1, \mathbb{E}[X], \mathbb{E}[X^{\otimes 2}], \mathbb{E}[X^{\otimes 3}], \ldots) \in \mathcal{H}$$

Using a well-known characterization of the mean as minimizer of a quadratic we can introduce the loss function

$$\mathcal{L}(x, a) := \|\varphi(X) - a\|^2,$$

where $\mathcal{A}$ is taken to be all of $\mathcal{H}$, which associates with a probability measure μ on $\mathbb{R}^n$ the Bayes action

$$a_\mu \equiv \mathbb{E}[\varphi(X)] \equiv \operatorname*{argmin}_{a \in \mathcal{H}} \mathbb{E}[\mathcal{L}(X, a)].$$

It follows from general principles that the resulting scoring rule on the state $\mathcal{X} = \mathbb{R}^n$ and action space $\mathcal{A} = \mathcal{H}$,

$$s(x, \mu) = \mathcal{L}(x, a_\mu)$$

is proper. Despite the elementary nature of this example it gives us a simple way to associate with any "feature map" φ a Bayes action and a scoring rule. Already simple variations lead to interesting research questions: for example, if the quadratic loss function is replaced by the absolute value, one ends with medians of moment as Bayes action and many other choices are possible. Such questions fall under the framework of "elicitation" of properties of probability measures with scoring rules which is an active research area, already in the classical vector-valued (even scalar) case; see [161] and [1, 178, 79] for some of the recent advances.

Remark 5.2.1. The question which probability measures on $\mathbb{R}^n$ are characterized by the moment sequence (5.2) is classical but quite subtle in general. But for compactly supported measures this trivially holds. Our focus will soon shift to probability

measures on pathspace but since spaces of paths are generically not even locally compact, compactness is a too strong assumption; in fact, important examples of measures on pathspace such as geometric Brownian motion are not characterized by their "signature moments" that we will use in Section 5.3 and Section 5.4. However, one can replace the moment sequence (5.2) by a normalized moment sequence that characterizes any probability measure on $\mathbb{R}^n$ and this extends to path space and signature moments, see [46] for details. Hence, for simplicity, we assume throughout that the probability measures are characterized by their expected feature map (since this is possible by a slight modification of the feature map).

5.3 Structure of the Space of (Unparametrized) Paths

The toy example in Section 5.2 uses the feature map $\varphi : \mathbb{R}^n \to \mathcal{H}$, equation (5.4), to construct scoring rules on $\mathbb{R}^n$. Our aim to follow the same approach by replacing $\mathbb{R}^n$ with the space of paths resp. tracks, hence we need a feature map defined on these domains. Seminal work of Chen [41] provides exactly such a generalization of φ; it shows the existence of a *"feature map"*

$$\Phi : \text{Tracks} \to \mathcal{H}, \quad \mathbf{x} \mapsto \Phi(\mathbf{x})$$

The co-domain $\mathcal{H}$ is the same as for φ. In the rest of this section we formally introduce the spaces Tracks and Paths, show how Φ faithfully represent elements of Tracks as elements of $\mathcal{H}$ and turns concatenation and reversal into algebraic operations in $\mathcal{H}$. Finally, we discuss why one can without loss of generality work on Tracks rather than Paths and in continuous time rather than discrete time.

Readers with less interest in the mathematical background might want to skip this section but have a brief look at Theorem 5.3.1 that formalizes how Φ represents tracks as elements of $\mathcal{H}$ and how it respects the operations of concatenation and reversal by turning them into the algebraic operations of multiplication and inversion.

The Domain of Paths. A bounded variation path[3] $\mathbf{x}$ in $\mathbb{R}^n$ is a continuous map

$$\mathbf{x} : [0, T] \to \mathbb{R}^n \text{ such that } \|\mathbf{x}\| := \sup_{(t_1,\dots,t_L):0\le t_1<\dots<t_L<T} \sum \|\mathbf{x}(t_{i+1}) - \mathbf{x}(t_i)\| < \infty.$$

[3]For simplicity we focus on (equivalence classes of) bounded variation paths but all the results immediately extend to paths with much less regularity such as trajectories of stochastic differential equations or (fractional) Brownian motion by replacing the iterated Riemann–Stieltjes integrals by stochastic integrals or rough path integrals [129]

For $a, b \in \mathbb{R}^n$ we denote with $\mathrm{Paths}(a, b)$ the set of all continuous bounded variation paths that start at a and end in b,

$$\mathrm{Paths}(a, b) := \{\mathbf{x} | \mathbf{x} : [0, T] \to \mathbb{R}^n \text{ is of bounded variation}, T > 0, \mathbf{x}(0) = a, \mathbf{x}(T) = b\}.$$

and by

$$\mathrm{Paths} := \bigcup_{a,b \in \mathbb{R}^n} \mathrm{Paths}(a, b)$$

the set of all bounded variation paths in $\mathbb{R}^n$. Although Paths is not a linear space, it has a rich structure given by concatenation and time reversal. Informally, this says that if one can go from a to b and from b to c then one can go from a to c and that if one can go from a to b then one can go from b to a. Formally, concatenation and reversal are defined as

1. For $\mathbf{x} \in \mathrm{Paths}(a, b)$, $\mathbf{y} \in \mathrm{Paths}(b, c)$, their *concatenation* $\mathbf{x} \star \mathbf{y} \in \mathrm{Paths}(a, c)$ is defined by

$$(\mathbf{x} \star \mathbf{y})(t) := \begin{cases} \mathbf{x}(t), \text{ if } t \in [a, b] \\ \mathbf{x}(b) - \mathbf{y}(b) + \mathbf{y}(t), \text{ if } t \in [b, b + c] \end{cases}$$

2. for any $\mathbf{x} \in \mathrm{Paths}(a, b)$ there exists an *inverse* path $\overleftarrow{\mathbf{x}} \in \mathrm{Paths}(b, a)$ defined as

$$\overleftarrow{\mathbf{x}}(t) := \mathbf{x}(b - t).$$

The Domain of Tracks As discussed above, often we want to ignore the time parametrization, hence the fundamental object we care about is not the set of paths but equivalence classes of paths. It turns out that is useful to work with slightly more general equivalence relation, namely that of *tree-like equivalence* $\sim$. We define

$$\mathrm{Tracks}(a, b) := \mathrm{Paths}(a, b)/\!\sim, \text{ and } \mathrm{Tracks} := \mathrm{Paths}\,/\!\sim$$

With slight abuse of notation, we use the same notation $\mathbf{x}$ for an element of Paths and an element of Tracks but emphasize that an element $\mathbf{x} \in \mathrm{Tracks}$ is a whole equivalence class of paths. We give the precise definition of the equivalence relation $\sim$ in Appendix 5.A and only note here that if two paths $\mathbf{x} : [0, T] \to \mathbb{R}^n, \mathbf{y} : [0, S] \to \mathbb{R}^n$ differ by time-parametrization, that is $\mathbf{x}(t) = \mathbf{y}(\varphi(t))$ for every t and an increasing function $\varphi : [0, T] \to [0, S]$, then $\mathbf{x} \sim \mathbf{y}$. However, in addition to time parametrization, tree-like equivalence also identifies paths that backtrack all their excursions, see Appendix 5.A. We invite readers to think of elements of Tracks like animal tracks in nature:

they provide shape and direction but not the speed at which the track was made. In particular, we note that the above operations of concatenation and reversal are well-defined for the elements of Tracks; after all, they do not depend on the time-parametrization. So again, with slight abuse of notation we have concatenation and reversal map,

$$\star : \mathrm{Tracks}(a,b) \times \mathrm{Tracks}(b,c) \to \mathrm{Tracks}(a,c) \text{ and } \overleftarrow{\bullet} : \mathrm{Tracks}(a,b) \to \mathrm{Tracks}(b,a).$$

The co-domain $\mathcal{H}$. Our first encounter of $\mathcal{H}$ was in Section 5.2 as the state space of the moment map (5.3). However, a more abstract way to introduce is by identifying it as the free algebra over $\mathbb{R}^n$. Informally, this means we want to keep the vector space structure of $\mathbb{R}^n$ but we also would like to have a multiplication and do this in the most general way possible. Formally, this means $\mathcal{H}$ is the free algebra over $\mathbb{R}^n$. Despite this abstract characterization as a free object, the space $\mathcal{H}_n$ has a very concrete form which we will take as its definition,

$$\mathcal{H} := \prod_{m \geq 0} (\mathbb{R}^n)^{\otimes m} := \{\mathbf{t} = (\mathbf{t}_0, \mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3, \ldots) : \mathbf{t}_m \in (\mathbb{R}^n)^{\otimes m}\}.$$

(one can then directly verify that this indeed is the free algebra, see [152]. That is, an element $\mathbf{t}$ of $\mathcal{H}$ is sequence of tensors $(\mathbf{t}_0, \mathbf{t}_1, \mathbf{t}_2, \ldots)$ of increasing degree m where by convention $(\mathbb{R}^n)^{\otimes 0} = \mathbb{R}$. The vector space structure of $\mathcal{H}$ is simply given as element-wise addition: addition of $\mathbf{s}, \mathbf{t} \in \mathcal{H}$ is defined as

$$\mathbf{s} + \mathbf{t} = (\mathbf{s}_0 + \mathbf{t}_0, \mathbf{s}_1 + \mathbf{t}_1, \ldots)$$

and their multiplication is defined by $(\mathbf{s} \cdot \mathbf{t})_m = \sum_k \mathbf{s}_k \otimes \mathbf{t}_{m-k}$, i.e

$$\mathbf{s} \cdot \mathbf{t} := (1, \mathbf{s}_1 + \mathbf{t}_1, \mathbf{s}_2 + \mathbf{s}_1 \otimes \mathbf{t}_1 + \mathbf{t}_2, \ldots)$$

where $\otimes$ denotes the usual tensor (outer) product. Like matrix multiplication, this multiplication is associative but in general not commutative, $\mathbf{s} \cdot \mathbf{t} \neq \mathbf{t} \cdot \mathbf{s}$ and it has as multiplicative unit $(1, 0, \ldots) \in \mathcal{H}$,

$$\mathbf{t} \cdot (1, 0, 0, \ldots) = (1, 0, \ldots) \cdot \mathbf{t} = \mathbf{t}.$$

The existence of a unit for multiplication naturally leads to the question of the existence of inverses, that is for $\mathbf{t} \in \mathcal{H}$ can one find another element in $\mathcal{H}$, denoted by $\mathbf{t}^{-1} \in \mathcal{H}$, such that

$$\mathbf{t} \cdot \mathbf{t}^{-1} = \mathbf{t}^{-1} \cdot \mathbf{t} = (1, 0, 0, \ldots).$$

This is true whenever $\mathbf{t}_0 \neq 0$, and moreover, $\mathbf{t} \mapsto \mathbf{t}^{-1}$ has the explicit formula

$$\mathbf{t}^{-1} = \frac{1}{\mathbf{t}_0}\Big\{ \sum_{m\geq 0} (1 - \frac{1}{\mathbf{t}_0}\mathbf{t})^{\otimes m} \Big\}.$$

For the rest of this paper, $\mathcal{H}$ will be equipped with its l^2 norm, so that if $\mathbf{t} = (\mathbf{t}_0, \mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3, \ldots) \in \mathcal{H}$, then

$$\|\mathbf{t}\|^2 = \|\mathbf{t}_0\|^2_{\mathbb{R}} + \|\mathbf{t}_1\|^2_{\mathbb{R}^n} + \|\mathbf{t}_2\|^2_{\mathbb{R}^n \otimes \mathbb{R}^n} + \cdots$$

Note that this makes $\mathcal{H}$ into a separable Hilbert space and a Banach algebra.

The Feature map Φ : Tracks $\to \mathcal{H}$.

Definition 40. For $\mathbf{x} \in$ Paths, $\mathbf{x} : [0, T] \to \mathbb{R}^n$ define

$$\int d\mathbf{x}^{\otimes m} := \int_{0\leq t_1 < \cdots < t_m \leq T} d\mathbf{x}(t_1) \otimes \cdots d\mathbf{x}(t_m) = \int \dot{\mathbf{x}}(t_1) \otimes \cdots \dot{\mathbf{x}}(t_m) dt_1 \cdots dt_m.$$

It is known that if two paths $\mathbf{x}, \mathbf{y} \in$ Paths are tree-like equivalent, $\mathbf{x} \sim \mathbf{y}$, then $\int d\mathbf{x}^{\otimes m} = \int d\mathbf{y}^{\otimes m}$ for every $m \geq 0$, see [97]. In fact, for the case of reparametrization $\mathbf{x}(t) = \mathbf{y}(\tau(t))$ this follows immediately from the change of variables formula. With slight abuse of notation we now define $\int d\mathbf{x}^{\otimes m}$ for $\mathbf{x} \in$ Tracks.

Definition 41. For $\mathbf{x} \in$ Tracks, define

$$\int d\mathbf{x}^{\otimes m} := \int d\mathbf{x}_{\text{path}}^{\otimes m}$$

where $\mathbf{x}_{\text{path}} \in$ Paths is in the equivalence class of $\mathbf{x}$ and $\int d\mathbf{x}_{\text{path}}^{\otimes m}$ is as in Definition 40.

By [97] $\int d\mathbf{x}^{\otimes m}$ for $\mathbf{x} \in$ Tracks is well-defined in the sense that the choice of $\mathbf{x}_{\text{path}}$ does not matter. We refer to the resulting map as the signature map (this is also known as the Chen–Fliess series or chronological exponential).

Definition 42. We call

$$\Phi : \text{Tracks} \to \mathcal{H}, \quad \mathbf{x} \mapsto \Big(\int d\mathbf{x}^{\otimes m} \Big)_{m\geq 0}$$

the signature map.

A well-known key property of the map Φ is that concatenation and reversal in Tracks correspond to multiplication and inversion in $\mathcal{H}$. Further, the map Φ is universal (up to fixing the starting point of the track, which is why we fix the starting point $a \in \mathbb{R}^n$ and restrict to the domain $\bigcup_{b\in\mathbb{R}^n} \text{Tracks}(a, b)$) in the sense that it linearizes continuous functions on Tracks. We summarize all this in Theorem 5.3.1 below.

Theorem 5.3.1. *For every $a \in \mathbb{R}^n$ the map*

$$\Phi : \bigcup_{b \in \mathbb{R}^n} \mathrm{Tracks}(a, b) \to \mathcal{H}, \quad \mathbf{x} \mapsto \left(\int d\mathbf{x}^{\otimes m} \right)_{m \geq 0}$$

is injective and

1. $\Phi(\mathbf{x} \star \mathbf{y}) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{y})$
2. $\Phi(\overleftarrow{\mathbf{x}}) = \Phi(\mathbf{x})^{-1}$
3. *for every $f \in C(\mathrm{Tracks}, \mathbb{R})$, $\epsilon > 0$ there exists a linear functional $\ell \in \mathcal{H}^\star$ such that*

$$|f(\mathbf{x}) - \langle \ell, \Phi(\mathbf{x}) \rangle| < \epsilon$$

uniformly in $\mathbf{x}$ on compacts.

Proof. This is a folk theorem in algebraic topology and control theory; see [41] and [71]. What is less standard is that we use the treelike equivalence from $\sim$ from [97]. □

Remark 5.3.1. The space $\mathcal{H} \equiv \prod_{m \geq 0} (\mathbb{R}^d)^{\otimes m}$ is graded by the tensor degree m, and $\Phi(\mathbf{x}) \equiv (\int dx^{\otimes m})_{m \geq 0}$ decays factorially in m, that is

$$\| \int d\mathbf{x}^{\otimes m} \| \leq \frac{\|\mathbf{x}\|^m}{m!}$$

(on the right hand side $\| \bullet \|$ denotes the bounded variation (semi-)norm, on the left-hand side it denotes the norm on $(\mathbb{R}^d)^{\otimes m}$). Hence, in practice one only needs to compute the first M iterated integrals of $\Phi(\mathbf{x})$. For piecewise linear tracks $\mathbf{x} \in \mathrm{Tracks}$ – which is how we identify time series – the first M entries of the map $\Phi(\mathbf{x})$ can be computed in $O(Ld^M)$ computational steps: if a track is given by piecewise linear segments $v_1, \ldots, v_L \in \mathbb{R}^n$ then

$$\left(\int d\mathbf{x}^{\otimes m} \right)_{m \in 0,1,\ldots,M} = \exp(v_1) \cdots \exp(v_L),$$

where $\exp(v) := (1, v, \frac{v^{\otimes 2}}{2!}, \ldots) \in \mathcal{H}$. Hence, for a low dimensional state space, the map $\Phi(\mathbf{x})$ can be approximately in time that scales linearly in the length of the path.

A Group $G \subset \mathcal{H}$. Our feature map Φ injects a track into the linear space $\mathcal{H}$. It turns out the domain of Φ has more structure, namely it takes values in subset $G \subset \mathcal{H}$ that forms a group G. In Theorem 5.3.1 we have seen that path inversion corresponds to an "inverse" in $\mathcal{H}$, formally $\Phi(\overleftarrow{\mathbf{x}}) = \Phi(\mathbf{x})^{-1}$, and this inverse operation is exactly the group inverse. Much more can be said about this group structure; however, since this only plays a role in some of our proofs, we give details in Appendix 5.D.

From Discrete Time to Continuous Time. This section has so far focused on paths [resp. tracks], that evolves [resp. equivalence classes of evolutions] in continuous time $\mathbf{x} : [0, T] \to \mathbb{R}^n$. However, in practice one typically has only access to a discrete time observations $\mathbf{x}(t_1), \dots, \mathbf{x}(t_L) \in \mathbb{R}^n$ along some grid $0 \leq t_1 < \cdots < t_L \leq T$, that is a time series. Any time series can be identified as the piecewise linear path, for example by piecewise linear interpolation

$$t \mapsto \mathbf{x}(t_i) + \frac{t_i - t}{t_{i+1} - t_i}(\mathbf{x}(t_{i+1}) - \mathbf{x}(t_i)) \text{ for } t \in [t_i, t_{i+1})$$

and hence also as an element of Tracks after forgetting about the parametrisation. In principle, other interpolations such as splines are possible but piecewise linear interpolation is arguably the most canonical without any prior knowledge about the underlying path regularity or behaviour in between time samples; we discuss this in more detail in Appendix 5.E where we also give a numerical experiment about the time discretization. Working in continuous time when the original data is discrete might look cumbersome and unnecessary at first sight but it has several advantages: Firstly, all time series are embedded into the same space Paths respectively Tracks, even if the sample grid $t_1 < \cdots t_L$ varies from time series to time series, which would not be the case if one identifies time series as vectors. Secondly, this ensures consistency in terms of high-frequency limits when the grid gets finer, that is $\sup |t^m_{i+1} - t^m_i| \to 0$ as $m \to \infty$ for a sequence $(t^m_i)_{i=1,\dots,L_m}$. Finally, many popular models are naturally formulated in continuous time rather than discrete time.

From Tracks to Paths. Our guiding philosophy is that the fundamental object is the set Tracks rather than the set Paths since the former allows us to ignore the time parametrization; since any continuous function $\tau : [0, T] \to [0, T']$ can be used to reparametrize a path $t \mapsto \mathbf{x}(t)$ to $t \mapsto \mathbf{x}(\tau(t))$, working with Tracks factors out a large class of invariances. Nevertheless, for certain applications the parametrization matters – at least to a certain degree. However, this can be easily addressed by adding time as an additional coordinate: to emphasize the dimension n of the state space $\mathbb{R}^n$ in which the paths evolve we write Paths_n (instead of just Paths that we used until now); similarly Tracks_n for the set of equivalence classes of Paths_n. Given $\mathbf{x} \in \text{Paths}_n$ we embed

$$\text{Paths}_n \hookrightarrow \text{Paths}_{n+1}, \quad \mathbf{x} \mapsto (t \mapsto (t, \mathbf{x}(t))) .$$

That is a path evolving in $\mathbb{R}^n$ is turned into a path in $\mathbb{R}^{n+1}$ by simply adding an additional coordinate that is time itself. This makes the parametrization part of the

"shape" of the trajectory which in turn is exactly the information that distinguishes tracks, hence

$$\text{Paths}_n \hookrightarrow \text{Tracks}_{n+1}\,.$$

This injection shows that any scoring rule for tracks induces a scoring rule for paths.

5.4 Scoring Rules For Tracks and Paths

Motivated by the toy example in Section 5.2 with the feature map φ for data in $\mathbb{R}^n$, we now follow the analogous reasoning on the space Tracks by using the feature map Φ. Recall that in the toy example in Section 5.3, a scoring rule on $\mathbb{R}^n$ was defined via the loss function $\|\varphi(X) - m)\|^2$. Section 5.3 showed that Φ takes values in a subset of $\mathcal{H}$ that forms a group. This motivates us to

replace the additive inverse in $\varphi(X) - m$ by the group inverse to get $\Phi(\mathbf{X})m^{-1}$

to define a loss function and subsequently a scoring rule on Tracks. Our first main result Theorem 5.2.1 shows that this is indeed the case and we arrive at a scoring rule that respects the natural operations on Tracks. Consequences of this result are Proposition 5.4.1 and Corollary 5.4.2 which show how the associated entropy is invariant to time-reversal and behaves under conditioning on the past.

A Scoring Rule for Tracks.

Definition 43. Let $L : \mathcal{H} \to \mathbb{R}$ be convex with a unique minimum at the unit $(1, 0, 0, \ldots)$ of G. Define the left loss function as

$$\mathcal{L}^{\text{left}} : \text{Tracks} \times \mathcal{H}^\star \to [0, \infty), \quad (\mathbf{x}, m) \mapsto L(m^{-1}\Phi(\mathbf{x})).$$

where we denote[4] $\mathcal{H}^\star := \{\mathbf{t} \in \mathcal{H} : \mathbf{t}_0 = 1\}$. Following step (1) from the scoring rule framework of Section 5.2, the left Bayes' act is defined as

$$a_\mu^{\text{left}} := \operatorname*{argmin}_{m \in \mathcal{H}^\star} \mathbb{E}[\mathcal{L}^{\text{left}}(m, \mathbf{X})].$$

Following step (2) yields the proper scoring rule

$$s^{\text{left}}(\mathbf{x}, \mu) := \mathbb{E}[\mathcal{L}^{\text{left}}(a_\mu^{\text{left}}, \mathbf{x})].$$

[4]The reason to introduce $\mathcal{H}^\star$ is that unlike $\mathcal{H}$, it is a group while also convex as a subset of $\mathcal{H}$ in addition to being topologically closed (unlike the set of invertible elements of $\mathcal{H}$). These turn out to be a very useful properties. The following inclusions hold $G \subseteq \mathcal{H}^\star \subseteq$ invertible elements of $\mathcal{H} \subseteq \mathcal{H}$.

Following step (3) yields the (generalised) entropy, divergence, and mutual information

$$\begin{aligned} H^{\text{left}}(\mu) &:= \mathbb{E}_{\mathbf{X}\sim\mu}\mathcal{L}^{\text{left}}(a_\mu^{\text{left}}, \mathbf{X}) \\ d^{\text{left}}(\nu,\mu) &:= \mathbb{E}_{\mathbf{X}\sim\nu}\Big[\mathcal{L}^{\text{left}}(a_\mu^{\text{left}}, \mathbf{X}) - \mathcal{L}^{\text{left}}(a_\nu^{\text{left}}, \mathbf{X})\Big] \\ I^{\text{left}}(\mu,\nu) &:= H(\mu) - \mathbb{E}_{U\sim\nu}[H(\mu|U)] \end{aligned}$$

on the output space $\mathcal{X} = \text{Tracks}$ and the action space $\mathcal{A} = \mathcal{H}^\star$. Analogously we define the right loss function $\mathcal{L}^{\text{right}}(\mathbf{x}, m) := L(\Phi(\mathbf{x})m^{-1})$, right Bayes act a_μ^{right} and right scoring rule $s^{\text{right}}(\mathbf{x}, \mu)$ as well as right entropy, divergence, and mutual information.

To sum up, Definition 43 provides us with a a scoring rule s^{left}, an entropy H^{left}, a divergence d^{left}, and mutual information I^{left} on the space of Tracks. Our main theoretical results, Theorem 5.4.1, Proposition 5.4.1, and Corollary 5.4.2 show that this scoring rule framework indeed respects the underlying structure of Tracks, in particular concatention and reversal.

Theorem 5.4.1. *Let the output space be* $\mathcal{X} = \text{Tracks}$*, the action space* $\mathcal{A} = \mathcal{H}^\star$ *and*

$$\mathcal{L}^{left} : \text{Tracks} \times \mathcal{H}^\star \to [0,\infty) \text{ resp. } \mathcal{L}^{right} : \text{Tracks} \times \mathcal{H}^\star \to [0,\infty)$$

the loss functions from Definition 43. The following properties hold

1. *If L is coercive, that is $L(\mathbf{t}) \to \infty$ whenever $\|\mathbf{t}\| \to \infty$, then for any Borel measure μ such that L is μ-integrable, both a_μ^{left} and a_μ^{right} exist. If L is strictly convex, then they are unique.*

2. *If $\mu = \delta_{\mathbf{x}}$ then $a_\mu^{right} = a_\mu^{left} = \Phi(\mathbf{x})$*

3. *The Bayes' acts satisfy*

$$\begin{aligned} a^{left}_{\nu\star\mu|\nu} &= \nu(\Phi) a^{left}_{\mu|\nu} \\ a^{right}_{\mu\star\nu|\nu} &= a^{right}_{\mu|\nu}\nu(\Phi) \end{aligned}$$

   where $\nu(\Phi)$ denotes the pushforward measure of ν under Φ and by $\mu\star\nu|\nu$ we denote the law of $\mathbf{X}\star\mathbf{Y}|\mathbf{Y}$ where $\text{Law}(\mathbf{X}) = \mu$ and $\text{Law}(\mathbf{Y}) = \nu$.

4. *There exists a linear map $\alpha : \mathcal{H} \to \mathcal{H}$ that is explicitly defined in Appendix 5.D such that if L satisfies $L(\mathbf{t}) = L(\alpha(\mathbf{t}))$, then*

$$a^{right}_{\overleftarrow{\mu}} = \alpha(a^{left}_\mu), \quad a^{left}_{\overleftarrow{\mu}} = \alpha(a^{right}_\mu)$$

   where $\overleftarrow{\mu}$ denotes the measure μ given by running samples from μ backwards in time[5]

[5]Formally, if $\mathbf{X} \sim \mu$ then $\overleftarrow{\mu}$ is defined as the law of time-reversed $\overleftarrow{\mathbf{X}}$.

Before we give a proof, let us informally describe the content of Theorem 5.4.1. Item 1 and item 2 simply guarantee that under general conditions, the Bayes acts exist and behave as one would expect. Item 3 shows that conditioning on the trajectory observed in the past, turns into an algebraic operation on the Bayes' act. This is not only encouraging from a purely mathematical point but we also use this below to prove that entropy, mutual information, and divergence behave well under operations on tracks. Finally, item 4 formalizes that the choice between s^{left} or s^{right} is only a matter of convention of whether we choose to run forwards or backwards in time (which we have in mathematics but apparently not in life) and that a well-defined algebraic operation α allows to switch between these two points of view.

Proof of Theorem 5.4.1. We give the proofs for the right Bayes' act as the proofs for the left Bayes' act is similar.

(1)

Fix some measure μ on $\mathcal{X}$ and define the map $\psi : G_n \to \mathbb{R}$ by

$$\psi(x) = \mathbb{E}_\mu L(\Phi(\mathbf{X})x).$$

We want to show that ψ is convex, coercive and lower semicontinuous on $(\mathcal{H}_n, \|\cdot\|)$, as this guarantees the existence of a minimiser. This is because the unit ball of $(\mathcal{H}_n, \|\cdot\|)$ is weakly compact, hence we could choose some weakly compact and convex set C such that $\psi(x) > M$ outside of C, and since ψ is lower semicontinuous it is also weakly lower semicontinuous, and therefore since it is convex it achieves a minimum on C which must be a global minimum. Note that its minimiser must be $(a_\mu^{\text{right}})^{-1}$. It follows that if ψ is strictly convex, then the minimiser is unique. Note that if L is (strictly) convex, then

$$\begin{aligned}\psi(\frac{1}{2}x + \frac{1}{2}y) &= \mathbb{E}_\mu L(\frac{1}{2}\Phi(\mathbf{X})x + \frac{1}{2}\Phi(\mathbf{X})y) \leq(<)\ \frac{1}{2}\mathbb{E}_\mu L(\Phi(\mathbf{X})x) + \frac{1}{2}\mathbb{E}_\mu L(\Phi(\mathbf{X})y) \\ &= \frac{1}{2}\psi(x) + \frac{1}{2}\psi(y)\end{aligned}$$

hence ψ is also (strictly) convex. Since $(\mathcal{H}, \|\cdot\|)$ is a Banach algebra, we know that for any $\mathbf{t}, \mathbf{s} \in \mathcal{H}$ $\|\mathbf{t} \cdot \mathbf{s}\| \leq \|\mathbf{t}\| \cdot \|\mathbf{s}\|$. By taking multiplicative inverses, this implies that

$$\|\mathbf{t} \cdot \mathbf{s}\| \geq \frac{1}{\|\mathbf{s}^{-1}\|}\|\mathbf{t}\|.$$

for any invertible element $\mathbf{s}$. As μ is Borel, and $\mathcal{H}_n$ is a Polish space, μ is a Radon measure by [24, Theorem 7.1.7] and we may choose a compact set C such that

$\mathcal{P}(\mathbf{X} \notin C) \leq \varepsilon$, and define M to be $\sup_{X \in C} \|\Phi(X)^{-1}\|$. Then on C, $\|\Phi(X)x\| \geq \frac{1}{M}\|x\|$, and since

$$\psi(x) = \mathbb{E}_\mu L(\Phi(\mathbf{X})x) \geq \mathbb{E}_{\mu|_C} L(\Phi(\mathbf{X})x)$$

and L is coercive, so is ψ.

To see that ψ is lower semicontinuous, note that for a sequence $x_k \to x$

$$\liminf_k \psi(x_k) = \liminf_k \mathbb{E}_\mu L(\Phi(\mathbf{X})x_k) \geq \mathbb{E}_\mu L(\Phi(\mathbf{X})x) = \psi(x)$$

by Fatous Lemma, the assertion follows.

(2) Since L is minimised at the unit it is clear that for $\mu = \delta_{\mathbf{x}}$, $a_\mu^{\text{left}} = a_\mu^{\text{right}} = \Phi(\mathbf{x})$ is optimal since $(a_\mu^{\text{left}})^{-1}\Phi(\mathbf{x}) = \Phi(\mathbf{x})(a_\mu^{\text{right}})^{-1} = 1$.

(3) For $\mathbf{Y} \sim \nu$ we have

$$\begin{aligned} a_{\mu \star \nu|\nu}^{\text{right}} :=& \operatorname*{argmin}_{m \in \mathrm{T}(\mathbb{R}^n)} \mathbb{E}_{\mathbf{X} \sim \mu}\Big[L(\Phi(\mathbf{X} \star \mathbf{Y})m^{-1})|\mathbf{Y}\Big] = \\ & \operatorname*{argmin}_{m \in \mathrm{T}(\mathbb{R}^n)} \mathbb{E}_{\mathbf{X} \sim \mu}\Big[L(\Phi(\mathbf{X})\Phi(\mathbf{Y})m^{-1})|\mathbf{Y}\Big] = \\ & \operatorname*{argmin}_{m\Phi(\mathbf{Y})^{-1} \in \mathrm{T}(\mathbb{R}^n)} \mathbb{E}_{\mathbf{X} \sim \mu}\Big[L(\Phi(\mathbf{X})m^{-1})|\mathbf{Y}\Big] := a_{\mu|\mathbf{Y}}^{\text{right}}\Phi(\mathbf{Y}). \end{aligned}$$

(4) If L satisfies $L(\mathbf{t}) = L(\alpha(\mathbf{t}))$, then

$$\begin{aligned} a_{\overleftarrow{\mu}}^{\text{right}} :=& \operatorname*{argmin}_{m \in \mathrm{T}(\mathbb{R}^n)} \mathbb{E}_{\mathbf{X} \sim \overleftarrow{\mu}} L(\Phi(\mathbf{X})m^{-1}) = \\ & \operatorname*{argmin}_{m \in \mathrm{T}(\mathbb{R}^n)} \mathbb{E}_{\mathbf{X} \sim \mu} L(\alpha(\Phi(\mathbf{X}))m^{-1}) = \\ & \operatorname*{argmin}_{m \in \mathrm{T}(\mathbb{R}^n)} \mathbb{E}_{\mathbf{X} \sim \mu} L(\alpha(m^{-1})\Phi(\mathbf{X})) := \alpha(a_\mu^{\text{left}}). \end{aligned}$$

□

Entropy, Divergence, and Mutual Information on the Space of Tracks. We now focus on the (generalized) entropy, divergence and mutual information for probability measures on tracks that results from Definition 43.

Proposition 5.4.1. *For any two probability measures μ and ν on* Tracks *it holds that*

1. $H^{left}(\mu \star \nu|\mu) = H^{left}(\nu|\mu)$ *and* $H^{right}(\mu \star \nu|\nu) = H^{right}(\mu|\nu)$
2. *If L satisfies $L(\mathbf{t}) = L(\alpha\mathbf{t})$, then*

$$\begin{aligned} H^{right}(\mu) &= H^{left}(\overleftarrow{\mu}), \\ d^{right}(\nu, \mu) &= d^{left}(\overleftarrow{\nu}, \overleftarrow{\mu}), \\ I^{right}(\mu, \nu) &= I^{left}(\overleftarrow{\mu}, \nu) \end{aligned}$$

Proof. (1)

$$\begin{aligned} H^{\text{left}}(\mu \star \nu|\mu) &= \mathbb{E}_{X\sim\mu\star\nu|\mu} L((a^{\text{left}}_{\mu\star\nu|\mu})^{-1}\Phi(X)) \\ &= \mathbb{E}_{X\sim\nu|\mu} L((a^{\text{left}}_{\nu|\mu})^{-1}\nu(\Phi)^{-1}\nu(\Phi)\Phi(X)) = H^{\text{left}}(\nu|\mu). \end{aligned}$$

(2)

$$\mathbb{E}_{X\sim\nu} L(\Phi(X)(a^{\text{right}}_{\mu})^{-1}) = \mathbb{E}_{X\sim\nu} L(\alpha(a^{\text{right}}_{\mu})^{-1}\alpha\Phi(X)) = \mathbb{E}_{X\sim\overleftarrow{\nu}} L((a^{\text{left}}_{\overleftarrow{\mu}})^{-1}\alpha\Phi(X)).$$

The other equalities follow. □

Corollary 5.4.2. *If L satisfies $L(\mathbf{t}) = L(\alpha(\mathbf{t}))$ and a measure μ is* reversible*, that is, μ and $\overleftarrow{\mu}$ are equal up to their starting distribution, then*

$$H^{right}(\mu) = H^{left}(\mu).$$

Remark 5.4.1. An alternative to using the group structure in Definition 43 of the Bayes act is to use that the group G is embedded into the linear space $\mathcal{H}$ and use this linear structure. That is, we define a Bayes act as $a_\mu := \text{argmin}_{m\in \mathrm{T}(\mathbb{R}^n)} \mathbb{E}_{\mathbf{X}\sim\mu}[L(\Phi(\mathbf{X}) - m)^2]$. It is easy to show that this gives a proper scoring rule and that $a_\mu = \mathbb{E}[\Phi(\mathbf{X})]$. However, this scoring rule relies on the embedding of the group into its ambient vector space and does not account for or respect the group structure. Moreover, the resulting divergence and entropy reduce to just the Euclidean distance and usual variance. The same remark extends to (signature) kernel based scoring, where linear methods are used in an RKHS; see the discussion about non-kernel based scoring in the introduction.

5.5 Gradient Descent on the Space of Tracks

Given a smooth function $f : \mathbb{R}^n \to \mathbb{R}$, the simplest update rule for gradient descent is

$$x_{i+1} = x_i - \eta\nabla f(x_i) \tag{5.5}$$

and under additional regularity of f, the resulting sequence $(x_i)_i \subset \mathbb{R}^n$ converges to a minimizer of f. Our interest lies minimizing functions $f : \text{Tracks} \to \mathbb{R}$ but since Tracks is not a linear space, classical gradient descent can not be applied. However, we have seen that Φ provides an isomorphism between the space of tracks and the subset G of $\mathcal{H}$ that forms a group (up to forgetting the starting point of the track)

$$\text{Tracks} \simeq G$$

Hence, the minimization problem of a function f of tracks can be re-formulated as a minimization problem of a function $F = F(g) := f \circ \Phi^{-1}(g)$ on the free group G. That is, the general problem we try to solve is to find

$$\operatorname{argmin}_{g \in G} F(g) \text{ for } F : G \to \mathbb{R}$$

for any F in class of sufficiently "smooth" real-valued functions on G.

There have been many attempts to generalize gradient descent to non-linear domains. Arguably the case of Riemannian manifolds [26] is the most well-developed among these. However, the group G does not come with a Riemannian structure (to wit, only a Sub-Riemannian structure [135]). We follow here a somewhat different approach based on seminal work of Pierre Pansu [143] that directly uses the group structure to define gradients. We show that this gradient in turn allows us to give a straightforward generalization of the gradient update rule (5.5) from $\mathbb{R}^n$ to G, so that the resulting sequence $(g_i)_i \subset G$ converges to the minimizer.

Pansu Gradients. The gradient $\nabla f(x)(\cdot)$ of a function

$$f : \mathbb{R}^n \to \mathbb{R}$$

at a point x is a linear functional of $\mathbb{R}^n$ that can be defined as the limit

$$\nabla f(x)(h) := \lim_{h \to 0} \frac{f(x + \lambda h) - f(x)}{\lambda}.$$

Identifying $\mathbb{R}^n$ as a the additive group $(\mathbb{R}^n, +)$ one can regard the difference quotient that appears in the limit as applying to the group operation to x and λh. Hence, if we have also have a generalization of the multiplication with the scalar λ, then the above difference quotient makes sense for other groups. To formalize scalar multiplication, it turns out that the right notion is that of a Carnot group: a Carnot group is a Lie group C that carries a left-invariant geodesic distance $\operatorname{dist} : C \times C \to [0, \infty)$ and for each $\lambda > 0$ a bijection

$$\delta_\lambda : C \to C \text{ such that } \operatorname{dist}(\delta_\lambda(g), \delta_\lambda(h)) = \lambda \operatorname{dist}(g, h).$$

However, our focus is on the free group G and here the scaling δ_λ by a scalar $\lambda > 0$ has an explicit form

Proposition 5.5.1. *The group $G \subset \mathcal{H}$ equipped with the geodesic distance and*

$$\delta : \lambda \times G \to G, \quad \delta_\lambda g = \delta_\lambda(1, g_1, g_2, g_3, \ldots) = (1, \lambda g_1, \lambda^2 g_2, \lambda^3 g_3, \ldots).$$

forms a limit of Carnot groups.

We now have all we need to define the (Pansu) gradient. We denote by $G^\star$ the topological dual of G.

Definition 44. Let $f : G \to \mathbb{R}$. We define $Df : G \to G^\star$ as

$$Df(g)h = \lim_{\lambda \downarrow 0} \frac{f(g\delta_\lambda h) - f(g)}{\lambda}$$

whenever this limit exists and call $Df(g)h$ the Pansu derivative of f at g in direction h. If f has a Pansu derivative for all $g \in G$ then we say that f is Pansu differentiable. Analogous we define the spaces $C^k(G, \mathbb{R})$ of k-times Pansu differentiable functions.

The Pansu gradient behaves very similar to the classic linear gradient. For example, for the proof of convergence of gradient descent on G we make use of the following "Taylor expansion".

Lemma 5.5.2. *If $f \in C^3(G, \mathbb{R})$, then*

$$f(gh) = f(g) + Df(g)h_1 + \frac{1}{2}D^2 f(g)h_2 + O(\|h_3\|)$$

Proof. Consider the function $A : \mathbb{R} \to \mathbb{R}$

$$A(\lambda) = f(g\delta_\lambda h)$$

then A is C^3 and by a (classical linear) Taylor expansion we may write

$$A(1) = A(0) + \dot{A}(0) + \frac{1}{2}\ddot{A}(0) + O(\|A^{(3)}(0)\|)$$

which translates into the asserted equation since $h^{\otimes n}$ is contained in h_n. □

Remark 5.5.1. A popular approach to differentiating functions of paths is to use a Frechet derivative as used in Malliavin calculus, i.e. one identifies the space of paths as a linear space, see [37]. However, the above Pansu derivative is of a very different nature and – by construction – respects the non-Euclidean structure of paths resp. tracks.

Gradient Descent on G. The idea of gradient descent to take a step in a direction that minimises f in a neighbourhood B_h. In our (Lie) group G, a natural choice is to choose some vector $v \in \mathbb{R}^n$ and use an exponential neighbourhood $ge^{\eta v}$ of $g \in G$ where exp denotes the exponential from Lie algebra to Lie group. Hence, the question becomes how choose v to minimise $f(ge^{\eta v})$. To do so, note that

$$f(ge^{\eta v}) = f(g\delta_\eta e^v)$$

whenever $v \in \mathbb{R}^n$. Now using Lemma 5.5.2 shows

$$f(g\delta_\eta e^v) = f(g) + Df(g)e_1^{\eta v} + \eta^2 = f(g) + \eta Df(g) \cdot v + \eta^2.$$

This suggest that the direction of steepest descent in the exponential neighbourhood is indeed given by the Pansu derivative (here we identify $G^\star$ with G as G is embedded into the Hilbert space $\mathcal{H}$).

$$v = -Df(g).$$

This leads to the following geometric descent rule

$$g_{i+1} = g_i e^{-\eta Df(g_i)}. \tag{5.6}$$

As for classic gradient descent, we need a notion of convexity to guarantee convergence to a minimum. Notions of convexity on the group G and their relation to the Pansu derivative are a well-studied topic and we refer to [55] for background.

Definition 45. We say that a function $f : G \to \mathbb{R}$ is geometrically convex if

$$f(g\delta_\lambda(g^{-1}h)) \leq (1-\lambda)f(g) + \lambda f(h)$$

for any $0 \leq \lambda \leq 1$.

As in the linear case, one can show that if $f : G \to \mathbb{R}$ is geometrically convex and C^2, then $D^2 f(g)$ is positive definite everywhere. However, $D^2 f$ is not symmetric in general unlike in the linear case. Putting everything together allows us to mimic the convergence proof of gradient descent in linear spaces which ultimately justifies the above informal derivation of the update rule.

Theorem 5.5.1.

1. *The geometric update rule is transitive, that is, one may go from any point in the group to any other point using updates of the form* (5.6).

2. *If f is geometrically convex and bounded from below with bounded second Pansu derivatives, then for η sufficiently small the sequence in Equation* (5.6) *converges to a minimum*

Proof. Item 1 follows from Chow's theorem [47] which states that G is generated by simple exponentials. For Item 2 let L be a bound on the derivatives of f. We may write

$$f(gh) \leq f(g) + Df(g)h_1 + \frac{1}{2}L\|h_2\|.$$

Hence, if $g_{n+1} = g_n e^{-\eta Df(g_n)}$, then

$$f(g_{n+1}) \leq f(g_n) - \eta\|Df(g_n)\|^2 + \frac{1}{2}\eta^2 L\|Df(g_n)\|^2 = f(g_n) + (\frac{L}{2}\eta^2 - \eta)\|Df(g_n)\|^2$$

which is smaller than $f(g_n)$ for η small enough whenever $Df(g_n) \neq 0$. Hence this is a strictly decreasing sequence and converges to a minimum. □

Theorem 5.5.1 gives the analogous convergence guarantees as regular gradient descent and is simple to implement.

5.6 Experiments

We now apply the theoretical result of the previous sections to inference tasks for sequential data. Our first experiment, Section 5.6.1, uses the divergence as a distance between sequential data and competitive baselines are available in the form of DTW distances. Our second experiment, Section 5.6.2 demonstrates how the mutual information coming from our scoring rule framework, can measure dependencies that would be hard to obtain with other approaches. Our third experiment, Section 5.6.3 revisits the classic topic of hypothesis testing and kernel methods provide here a strong baseline.

Implementation Details. In all our experiments we take as loss function L the squared norm,

$$L(\mathbf{t}) = \|\mathbf{t}\|^2 = \sum_{m=1}^{M} \|\mathbf{t}_m\|^2.$$

which is symmetric by Lemma 5.D.1. Note that the function L is smooth as a sum of squares, and since Φ is also smooth, since it is polynomial in the increments of its input when computed up to some fixed degree M, we can easily compute d, H, and I by automatic differentiation as we may quickly find a_μ^{left} via gradient descent. For the computation of Φ we use the signatory [111] package which allows for fast and easy computations. We will use two different time warpings $\varphi_p(t)$ and $\psi_{c,M}(t)$ in these

experiments, both are designed to map the interval $[0, 1]$ to itself monotonically and are defined as follows

$$\varphi_p(t) = t^p, \quad \psi_{c,M}(t) = \begin{cases} \frac{t}{c}, & \text{if } t < M \\ \frac{M}{c} + (1 - \frac{M}{c}) \times \frac{t-M}{1-M}, & \text{otherwise} \end{cases}.$$

φ_p works by running time slowly in the beginning and quickly towards the end, the parameter $p \in [1, \infty)$ determines the severity of the warping. $\psi_{c,M}$ works by running time at two different speeds depending on if $t < M$ or not, the parameter $M \in [0, 1]$ determines where the change point lies, and $c \in (0, \infty)$ determines how big the change is. Both of these act on the unit interval, but can be used for any interval $[0, T]$ by considering $T\varphi_p(t/T)$ and $T\psi_{c,M}(t/T)$.

5.6.1 Experiment: Distances for Time-Series

One natural question to ask is how well the regularised versions of DTW that allow for differentiation are able to incorporate the parametrisation invariance. One drawback of these regularisation of DTW is that it typically leads to a trade-off between the smoothness and parametrization invariance, see Appendix 5.B for more on DTW and sDTW. Given two time series, $\mathbf{x}$ and $\mathbf{y}$, we then use sDTW $d_{\text{sDTW}}(\mathbf{x}, \mathbf{y})$ which is the regularised version of DTW introduced in [22] which is differentiable and to a certain degree parametrisation invariant quantity. The scoring rule framework that we presented in the previous sections provides divergence not only between time series but probability measures on time series,

$$d(\nu, \mu) \equiv \mathbb{E}_{X \sim \nu}[s(X, \mu)] - H(\nu),$$

but as a special case, we can restrict this to point measures to get a "divergence" between two time series akin to sDTW which reduces to the formula

$$(\mathbf{x}, \mathbf{y}) \mapsto d(\delta_{\mathbf{x}}, \delta_{\mathbf{y}}) = L(\Phi(\mathbf{x})\Phi(\mathbf{y})^{-1}) \tag{5.7}$$

which we call the *geometric score*, in this experiment we truncate Φ at level 6. To study the different behaviour between these two "divergences" – sDTW and (5.7) – we generated samples paths from a Brownian motion to get $\mathbf{x}$ and set $\mathbf{y} = \mathbf{x}_{\varphi_p(t)}$. The results are shown in Figure 5.2, here p is a parameter that determines the severity of the warping, see Appendix 5.E. The geometric score stays close to 0 regardless of the value of p but the sDTW divergence will increase with p and the rate of increase depends on the γ parameter. As γ tends to 0 the sDTW score will tend to the geometric score, but it will lose its smoothness while doing so. It is worth noting that

$\gamma = 1$ is considered a default value. To achieve the same level of invariance enjoyed by the divergence (5.7) in a DTW paradigm on can use the classical non-smooth DTW algorithms which cannot be updated using gradient descent. In contrast, the signature divergence (5.7) is always differentiable and parametrization invariant. Further, the signature divergence is more general in the sense that it is not just a divergence between time series but between probability measures on time series which allows in principle for many other applications such as variational inference.

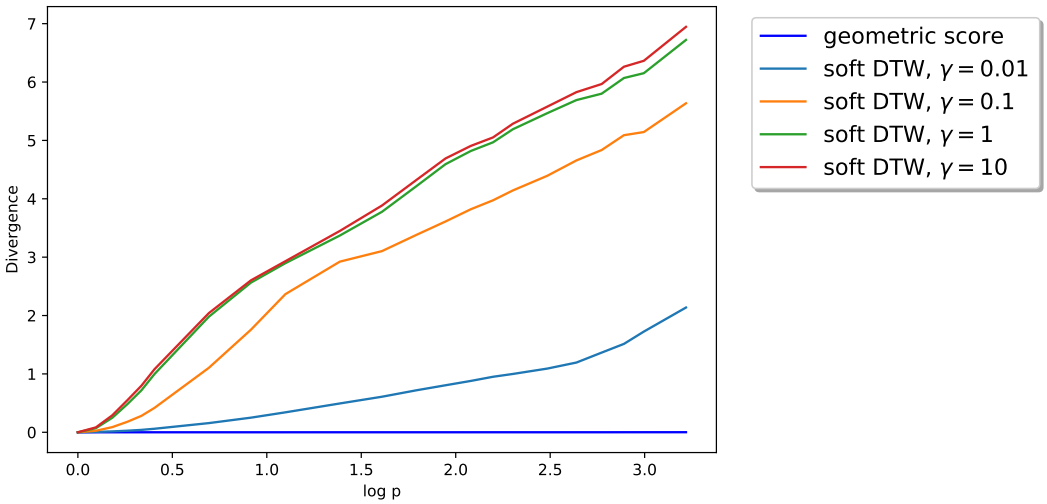


Figure 5.2: Geometric scoring compared to sDTW for different values of γ plotted against the logarithm of p for p between 1 and 25.

The implementation of sDTW is taken from the excellent Python package from [22].

5.6.2 Experiment: Non-linear Dependencies in Time Series

Recall that the mutual information between two probability measures μ and ν is defined as

$$I(\mu, \nu) := H(\mu) - \mathbb{E}_{U \sim \nu}[H(\mu|U)]. \tag{5.8}$$

and provides a dependency measure between μ and ν. Closest related to independence on paths are the *signature cumulants* [30] which have been proven to be useful in applications [163, 75]. However signature cumulant only compare probability measures on paths [tracks] with other probability measures on paths [tracks]. In contrast, the mutual information (5.8) allows to measures dependency between a probability measure μ on paths [tracks] and a probability measure ν on an arbitrary measurable

space; in particular, this allows to measure dependence relations between time series and scalar-valued random variables. Equation 5.8 can be efficiently computed whenever one is able to sample from the conditional distribution $\mu|U$.

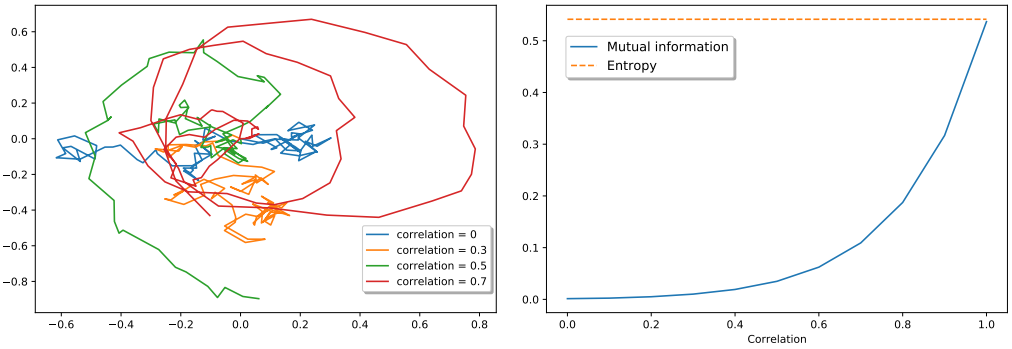


Figure 5.3: The left hand side shows sample trajectories from $\mathbf{Y}$ for different correlation values ρ. The right plot shows the mutual information between $\mathbf{Y}$ and the rotational speed ω as a function of ρ.

To demonstrate this, we consider two examples, the mutual information is computed by first sampling from the second random variable, and then sampling from the conditional measure of the first random variable.

Mutual information between a stochastic process and a scalar. We consider the stochastic process $\mathbf{y}$ defined as

$$\mathbf{y}_t = \rho t \begin{pmatrix} \cos \omega t \\ \sin \omega t \end{pmatrix} + \sqrt{1 - \rho^2} \mathbf{x}_t$$

where ω is sampled from the uniform distribution on $[0, 8\pi)$ and $\mathbf{x}$ is a standard 2-dimensional Brownian motion independent from ω. $\mathbf{x}$ and $\mathbf{y}$ are generated on the interval $[0, 1]$. The left plot in Figure 5.3 shows some sample trajectories of the process $\mathbf{y}$. The right plot in Figure 5.3 shows the mutual information between (the law of) ω and $\mathbf{y}$ for various values of ρ. As expected, the mutual information increases monotone with ρ and is bounded by the entropy.

Mutual Information between two unparametrized stochastic processes. We consider two independent a Brownian motions $\mathbf{x}$ and $\mathbf{y}$ as before, and a random parametrisation $\varphi_p(t)$ where $p \in [1, 10]$ is uniformly distributed independent of $\mathbf{x}$. The stochastic process $\mathbf{z}$ is defined as

$$\mathbf{z}_t = \rho \mathbf{x}_{\varphi_p(t)} + \sqrt{1 - \rho^2} \mathbf{y}_t.$$

We then compute the mutual information between $\mathbf{z}$ and $\mathbf{x}$ as a function of ρ. The results are shown in Figure 5.4.

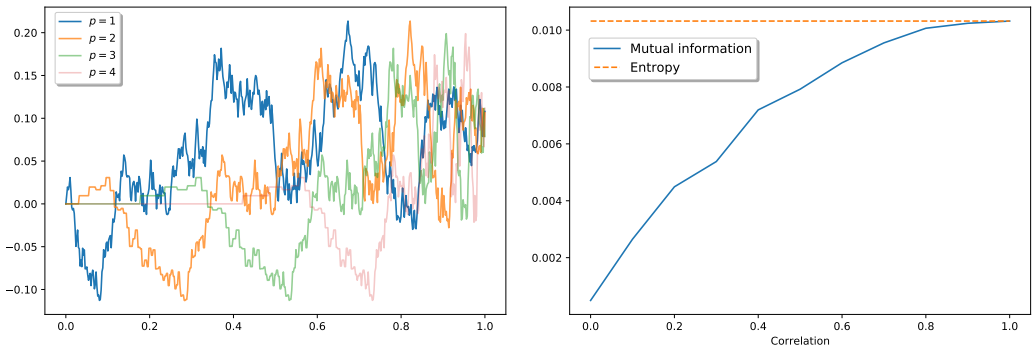


Figure 5.4: The left plot shows one-dimensional projections of sample paths of $\mathbf{x}$ for different warping parameters p. The right plot shows the mutual information between $\mathbf{x}$ and $\mathbf{z}$ as a function of ρ.

5.6.3 Experiment: Hypothesis Testing

We consider the following two testing problems

$$H_0 : \mu = \nu \text{ against the alternative } H_1 : \mu \neq \nu$$
$$H_0 : \gamma = \mu \otimes \nu \text{ against the alternative } H_1 : \gamma \neq \mu \otimes \nu.$$

The former is known as two-sample test, the latter as independence test. Given a finite number of samples and a test statistic, a permutation test provides a simple and robust test method; we refer to [96] for general background. Below we compare the performance of a permutation test with statistics from our scoring rule framework (divergence and mutual information) with widely used statistics from kernel learning (MMD and HSIC).

Two-Sample Testing. We consider the two-sample test

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

on the space Tracks; that is, based time-series sampled from two-probability measures μ and ν we need to decide whether the measures μ and ν produce samples that only differ by reparametrization. To carry out a two-sample test, we need statistics that quantify similarity between probability measures. Given a kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, the associated *MMD distance*

$$\mathrm{MMD}(\mu, \nu) = \|\mathbb{E}_{X\sim\mu} k(\cdot, X) - \mathbb{E}_{X\sim\nu} k(\cdot, Y)\|_{\mathcal{H}_k}$$

provides such a quantitity that can be estimated from finite samples from μ and ν, see Appendix 5.C for background. We use the arguably most popular choice for k, namely the RBF kernel [35] and follow the standard machine pipeline to identify a time-series as a vector: a real-valued time series $z(t_1), \ldots, z(t_L)$ is identifed as an element of $\mathbb{R}^d$ for $d = L$. Note that the resulting test statistic requires to compute a $(L \times L)$-Gram matrix which quickly becomes costly for time series sampled with high-frequency resp. large L. From our scoring rule framework, the divergence

$$d^{\text{right}}(\nu, \mu) = \mathbb{E}_{X \sim \nu} s^{\text{right}}(\mu, X) - H^{\text{right}}(\nu).$$

provides another quantity that quantities the difference and that can be simply estimated by a plug-in estimator. By construction we expect d^{right} to be invariant to the paremetrization.

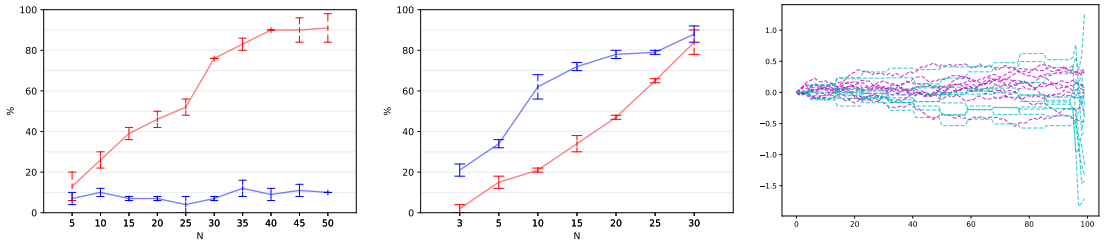
We use both of these statistics, to carry out test of level α of 5%. We repeat the permutation tests for different sample regimes. The two types of errors associated to these tests are type I and type II errors[6].

Type I. We simulate two independent Brownian motions $\mathbf{x}, \mathbf{z}$ and set $\mathbf{y} = \mathbf{z}_{\varphi_p(t)}$ where p is uniformly distributed on $[1, 5]$. These two measures are the same up to a time change. As expected, the kernel method fails to recognize this invariance up to a time change and falsely rejects the null as the samples increase; in contrast, the divergence classifier hovers around $5 - 10\%$ rejection rate, see Figure 5.5a.

Type II. We simulated independent Brownian paths $\mathbf{x}, \mathbf{z}$ and set $\mathbf{y} = \sqrt{c}\mathbf{z}$ scaled to have variance ct at time t. We then considered the paths $\mathbf{y}_{\psi_{c,M}(t)}$, so that $\mathbf{y}$ is ran slower for most of the time, see Figure 5.5c for sample paths of $\mathbf{x}, \mathbf{y}$. We used $c = 9$ and $M = 0.95$ in the experiments. The accuracies of the MMD classifier and the divergence classifier are shown in Figure 5.5b. They both go up to 100% as the number of samples increase, but the divergence classifier performs significantly better for lower sample sizes.

The results are shown in Figure 5.5. For every sample size N we ran each experiment several times and the accuracies of both runs are shown as intervals, along with their mean as a line plot in Figures 5.5a,5.5b. Similarly, the type II experiments are shown in Figure 5.5c.

[6]Type I error is when you falsely conclude that two of the same measures are different, by design this error rate should be the same as your α, so 5% in our case. Type II error is when you fail to say that two different measures are different.



(a) Type I accuracy, it should be around 5%.

(b) Type II accuracy, it should be as high as possible.

(c) Parametrised sample paths from the first marginal of the two compared measures in 5.5b.

Figure 5.5: The accuracy (%) of the permutation classifiers as a function of the sample size (N). The kernel MMD classifier is shown in red and the divergence classifier d^{right} is shown in blue.

Independence Testing. We consider the hypothesis testing problem

$$H_0 : \gamma = \mu \otimes \nu \text{ against the alternative } H_1 : \gamma \neq \mu \otimes \nu.$$

where γ is a measure on the product space $\mathbb{R} \times \text{Tracks}$ and μ is the marginal distribution of γ on $\mathbb{R}$ and ν the marginal distribution on Tracks. The example we consider is a model for time series (the measure ν) that is parametrized by a real number (the measure ν). The independence test asks whether the dependence on the parameter happens only via the parametrization This is a subtle question since the correlation might not be instantaneous, i.e. the parameter can influence the whole shape of the trajectory. Our benchmark is again the kernel learning approach which uses the MMD distance

$$\text{MMD}(\gamma, \mu \otimes \nu).$$

This is known as the *Hilbert-Schmidt independence criterion* (HSIC); see [94] for background. Our scoring rule framwork provides another quantitiy, namely the mutual information arising from our scoring rule,

$$I(\mu, \nu) = H(\mu) - \mathbb{E}_{U \sim \nu}[H(\mu|U)].$$

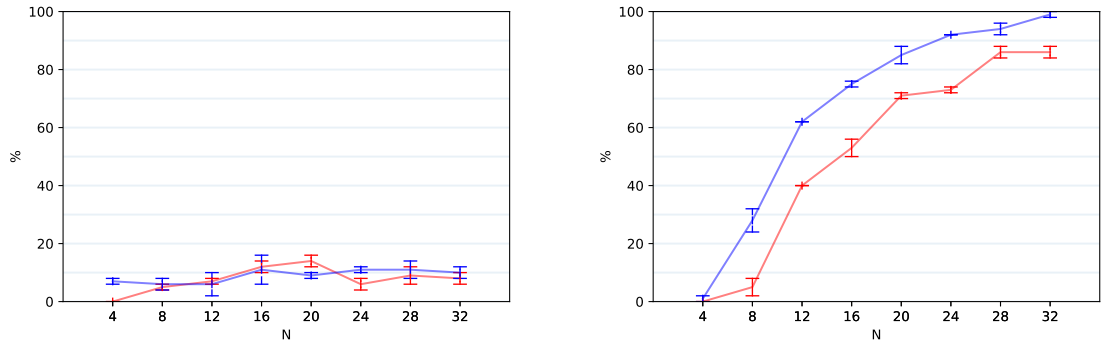
We again use permutation test and compare the use of HSIC and mutual information as test statistic.

Type II We generate the same paths as in Section 5.6.2, that is, we set

$$\mathbf{z}_t = \rho \mathbf{x}_{\varphi_p(t)} + \sqrt{1 - \rho^2} \mathbf{y}_t.$$

where $\mathbf{x}, \mathbf{y}$ are independent Brownian motions, $\rho = 0.7$, and p is generated uniformly on $[1, 10]$. We compare the distributions of $\mathbf{y}$ and $\mathbf{z}$, and sample from the conditional measure by generating 4 values of $\mathbf{z}$ for every 1 value of $\mathbf{y}$. We tested this on a sample size from 4 to 32, see Figure 5.6b for the results.

Type I We compare two independent Brownian motions, see Figure 5.6a for the results.



(a) Type I accuracy, it should be around 5%. (b) Type II accuracy, it should be as high as possible.

Figure 5.6: The accuracy (%) of the permutation classifiers as a function of the sample size (N). The kernel HSIC classifier is shown in red and the mutual information classifier I is shown in blue.

Computational Complexity. We only discussed the performance of the different test statistics but they also differ in computational complexity. To compute the scoring rule and associated quantities (divergence, mutual information, etc) requires an initial computation of the law of $\Phi(x)$ and then finding the Bayes' act using gradient descent. The former is an infinite series of tensors, $\Phi(x) \in \prod_{m \geq 0} (\mathbb{R}^d)^{\otimes m}$ but decays factorially fast, Remark 5.3.1, hence one can truncate after the first m tensors; in practice $m \approx 7$ is more than sufficient. Computing the first m terms $(1, \int d\mathbf{x}, \dots \int d\mathbf{x}^{\otimes m})$ for a discrete time series $x(t_1), \dots, x(t_L) \in \mathbb{R}^d$ can be done in time linear in L, and d^m memory since an element of $(\mathbb{R}^d)^{\otimes m}$ has d^m coordinates. Finding the Bayes' act via gradient descent scales linearly in the number of samples (one sample is a whole time series). To sum up, the scoring rule approach scales linearly in the number of samples and linearly in the maximal number of time points L but $O(d^m)$ in the dimension of the state space. In contrast, kernel MMD or HSIC requires the computation of a Gram matrix which scales typically linearly in state space dimension d but quadratically

in the number of samples. Thus the bottleneck for scoring rule is the dimension of the state space, whereas the bottleneck for kernel methods is the sample size. In our experiments, neither was prohibitive and all experiments were carried out on a laptop.

5.7 Conclusion

We developed a scoring rule framework that applies to sequential data in discrete and continuous time. Our approach focused on respecting the non-linear structure of sequential data which we accomplish by using ideas from pure mathematics. Despite the more abstract derivation, the resulting scoring rule, entropy, divergence, and mutual information are easy to compute and use in practice without knowledge of the algebraic background. Our experiments indicate favourable performance with standard approaches and further allow for novel application such as correlations with scalars and stochastic processes.

5.A Tree-like equivalence of paths

Informally, a path $x : [0, T] \to \mathbb{R}^n$ is *tree-like* if the set it traces out in $\mathbb{R}^n$ looks like a tree. The formal definition is

Definition 46 ([97])**.** A continuous path $x : [0, T] \to E$ is said to be *tree-like* if there exists an $\mathbb{R}$-tree τ, a continuous map $\varphi : [0, T] \to \tau$ and a map $\psi : \tau \to E$ such that $\varphi(0) = \varphi(T)$ and $x = \psi \circ \varphi$. Two paths $\mathbf{x}, \mathbf{y}$ are *tree-like equivalent* if $\mathbf{x} \star \mathbf{y}^{-1}$ is tree-like and we denote this relation with $\mathbf{x} \sim \mathbf{y}$.

That is, if $\mathbf{x}$ and $\mathbf{y}$ follow the same trajectory in $\mathbb{R}^n$ up to tree-like excursions then $\mathbf{x} \sim \mathbf{y}$. In particular, if $\mathbf{y}$ is just bx under time-reparametrization, $\mathbf{y}(t) = \mathbf{x}(\tau(t))$ for an increasing τ, then $\mathbf{x} \sim \mathbf{y}$. See also [30, Appendix B] for more details. It is easy to check that $\sim$ is an equivalence relation, hence we can define

$$\text{Tracks}(a, b) := \text{Paths}(a, b)/\!\sim, \text{ and } \text{Tracks} := \text{Paths}\,/\!\sim,$$

5.B Dynamic time warping

Given two time series $\mathbf{x} = (\mathbf{x}_i)_{i=1}^n, \mathbf{y} = (\mathbf{y}_j)_{j=1}^m$, let $\Delta(\mathbf{x}, \mathbf{y})$ denote their $n \times m$ distance matrix $\Delta(\mathbf{x}, \mathbf{y})_{i,j} := \|\mathbf{x}_i - \mathbf{y}_j\|$. An *alignment matrix* $A \in \{0, 1\}^{n \times m}$ is any $n \times m$ matrix with entries in $0, 1$ such that the 1's form a path from upper left to the lower right corner only moving down, right, or diagonally down and right, for example

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

is a valid alignment matrix, we denote the set of all $n \times m$ alignment matrices by $\mathcal{A}_{n,m}$. The cost associated to a given alignment matrix A can be computed as $\langle A, \Delta(\mathbf{x}, \mathbf{y})\rangle$. For example, with A as in Equation 5.B we get

$$\langle A, \Delta(\mathbf{x}, \mathbf{y})\rangle = \|\mathbf{x}_1 - \mathbf{y}_1\| + \|\mathbf{x}_2 - \mathbf{y}_2\| + \|\mathbf{x}_2 - \mathbf{y}_3\| + \|\mathbf{x}_3 - \mathbf{y}_3\| + \|\mathbf{x}_3 - \mathbf{y}_4\| + \|\mathbf{x}_3 - \mathbf{y}_5\|.$$

The dynamic time warping distance between $\mathbf{x}, \mathbf{y}$ is defined as

$$\text{DTW}(\mathbf{x}, \mathbf{y}) := \min_{A \in \mathcal{A}_{n,m}} \langle A, \Delta(\mathbf{x}, \mathbf{y})\rangle.$$

By design this is clearly invariant to the specific parametrisation of $\mathbf{x}$ and $\mathbf{y}$, but its main drawback, as stated in the main text, are that because of the need to compute the cost for every alignment, of which there are exponentially many (the number of

alignments of n and m is known as the delannoy number which grows exponentially), the computation is slow. One may speed this up by considering only alignments that are close to the identity which will be optimal or close to optimal in most cases but the complexity is still $O(nm)$ which makes it infeasible for long, or high-frequency data. Another drawback is that because it involves the min function it is not smooth, which makes it unsuitable for deep learning. The proposed remedy for its non-smoothness is to replace the min function by the softmin_γ function, defined as

$$\text{softmin}_\gamma(a_1, \ldots, a_n) = -\gamma \log \sum_{i=1}^{n} e^{-a_i/\gamma},$$

this version of DTW is known as soft DTW or sDTW. Note that

$$\text{as } \gamma \to 0, \quad \text{softmin}_\gamma(a_1, \ldots, a_n) \to \min(a_1, \ldots, a_n)$$

so γ can be thought of as a regularising parameter that introduces a trade off between smoothness and fidelity. Note that even if sDTW is smooth, it is still unsuitable for time series with many samples because of its quadratic complexity in the lengths n and m.

5.C Kernel methods

Kernel methods are traditionally very useful for high-dimensional data such as time-series. The idea is that one can leverage a feature map $\varphi : \mathcal{X} \to \mathcal{H}$ from your original space $\mathcal{X}$ to a high or infinite dimensional Hilbert space $\mathcal{H}$. When $\mathcal{H}$ is a so-called *reproducing kernel Hilbert space*, then one can represent inner products $\langle \varphi(x), \varphi(y) \rangle$ by a kernel function $k(x, y)$ with some very desirable properties, see [164] for more on kernels and kernel learning. The kernel used for the experiments in this paper was the rbf kernel:

$$k_{\text{rbf}}(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma}}$$

The parameter σ is known as the bandwidth for the kernel. We use both the *MMD* and the *HSIC* for two-sample testing and independence testing. Given two measures μ and ν, the MMD is defined as follows

$$\text{MMD}_k(\mu, \nu) = \|\mathbb{E}_{X\sim\mu} k(X, \cdot) - \mathbb{E}_{Y\sim\nu} k(Y, \cdot)\|_{\mathcal{H}_k},$$

and given a product measure γ with marginals μ, ν and two different kernels k_1, k_2, the HSIC is defined as:

$$\text{HSIC}_{k_1 \otimes k_2}(\gamma) = \text{MMD}_{k_1 \otimes k_2}(\gamma, \mu \otimes \nu).$$

Thes MMD can be computed by estimating the following expressions

$$\mathrm{MMD}_k(\mu,\nu) = \mathbb{E}_{\mu\otimes\mu}k(X,X') + \mathbb{E}_{\nu\otimes\nu}k(Y,Y') - 2\mathbb{E}_{\mu\otimes\nu}k(X,Y)$$

where X, X' denote independent copies of $X \sim \mu$ and Y, Y' denote independent copies of $Y \sim \nu$. The HSIC can be computed as

$$\begin{aligned}\mathrm{HSIC}_k(\gamma) = \mathbb{E}_{\gamma\otimes\gamma}k_1(X,X')k_2(Y,Y') + \mathbb{E}_{\mu\otimes\mu\otimes\nu\otimes\nu}k_1(X,X')k_2(Y,Y') \\ - 2\mathbb{E}_{\gamma\otimes\mu\otimes\nu}k_1(X,X')k_2(Y,Y').\end{aligned}$$

In this paper the MMD distance was computed using its U-statistic as this is the easiest and most common way to compute it [92], given two sets of samples $(x_i)_{i=1}^n, (y_i)_{i=1}^m$ and a kernel k, one estimates the MMD with the expression

$$\begin{aligned}\frac{1}{n(n-1)}\sum_{1\le i\le n}\sum_{1\le j\le n, j\neq i}k(x_i,x_j) + \frac{1}{m(m-1)}\sum_{1\le i\le m}\sum_{1\le j\le m, j\neq i}k(y_i,y_j) - \\ \frac{2}{nm}\sum_{1\le i\le n}\sum_{1\le j\le m}k(x_i,y_j).\end{aligned}$$

HSIC was computed using its V-statistic as it is faster and conceptually easier [94], given samples $(x_i)_{i=1}^n, (y_i)_{i=1}^n$ and two kernels k_1, k_2, define the matrices $K_1^{i,j} = k_1(x_i,x_j), K_2^{i,j} = k_2(y_i,y_j)$ and let C be the centering matrix $C = I - \frac{1}{n}11^T$. We can estimate HSIC with the expression

$$\mathrm{Tr}\Big(K_1CK_2C\Big).$$

Note that the statistics for HSIC and MMD both have quadratic complexity in the sample size.

5.D Hopf Algebras

The so-called *antipode* map

$$\alpha : \mathcal{H} \to \mathcal{H}$$

is defined as the linear function given by linear extensions of the map

$$(\mathbb{R}^n)^{\otimes m} \to (\mathbb{R}^n)^{\otimes m}, \quad v_1 \otimes \cdots \otimes v_m \mapsto (-1)^m v_m \otimes \cdots \otimes v_1.$$

There is an important subset G of $\mathcal{H}$ defined by the property

$$G := \{g \in \mathcal{H} : \alpha(g) = g^{-1}\}.$$

It turns out that G in fact forms a group and will play an important role in our Bayes acts for the simple fact that the feature map Φ takes values in G. We summaries this along with some facts about α that we use later in the following lemma.

Lemma 5.D.1. *Let α be the antipode on $\mathcal{H}$. Then*

1. $\alpha^2 = 1$,
2. $\alpha(\mathbf{s} \cdot \mathbf{t}) = \alpha(\mathbf{t})\alpha(\mathbf{s})$,
3. *If* $\mathbf{x} \in$ Tracks, *then* $\Phi(\mathbf{x}) \in G$,
4. *For a power series* $p(\mathbf{t}) = \sum_{m \geq 0} p_m \mathbf{t}^{\otimes m}$, *it holds that* $p \circ \alpha(\mathbf{t}) = \alpha \circ p(\mathbf{t})$ *for any* $\mathbf{t} \in \mathcal{H}$,
5. *If* $\mathbf{t} \in \mathcal{H}$ *is invertible, then* $\alpha(\mathbf{t}^{-1}) = \alpha(\mathbf{t})^{-1}$,
6. *Let* $f : \mathcal{H} \to \mathbb{R}$ *have the form*

$$f(\mathbf{t}) = \sum_m \sum_{i_1 \cdots i_m} f_m(\mathbf{t}_m^{i_1 \cdots i_m})$$

where we identify the degree-m tensor $\mathbf{t}_m \in (\mathbb{R}^d)^{\otimes m}$ *of* $\mathbf{t} = (\mathbf{t}_m)_{m \geq 0} \in \mathcal{H}$ *with its coordinates* $\mathbf{t}_m \simeq (\mathbf{t}_m^{i_1, \ldots, i_m})_{i_1, \ldots, i_m = 1, \ldots, d}$. *If* $f_m(x) = f_m(-x)$ *for every* $m \geq 0$ *and* $x \in \mathbb{R}$, *then*

$$f(\mathbf{t}) = f(\alpha(\mathbf{t})).$$

Proof. Items 1 and 2 follow from the definition. Item 3 is well known, see for instance [44, Section 5]. To see item 4 note that

$$p \circ \alpha(x) = \sum_{n \geq 0} p_n (\alpha x)^{\otimes n} = \sum_{n \geq 0} p_n \alpha(x^{\otimes n}) = \alpha(\sum_{n \geq 0} p_n x^{\otimes n}) = \alpha \circ p(x),$$

by Item 2 and linearity. Item 5 follows since the inverse map $\mathbf{t} \mapsto \mathbf{t}^{-1}$ has the power series expansion

$$\mathbf{t}^{-1} = \frac{1}{\mathbf{t}_0} \Big\{ \sum_{n \geq 0} (1 - \frac{1}{\mathbf{t}_0} \mathbf{t})^{\otimes n} \Big\},$$

see [78, Lemma 7.16] which together with Item 4 shows the claim. For item 6, we note that

$$f(\alpha x) = \sum_n \sum_{i_1 \cdots i_n} f_n((-1)^n x^{i_n \cdots i_1}) = \sum_n \sum_{i_1 \cdots i_n} f_n(x^{i_n \cdots i_1}) = f(x).$$

□

A simple example of a function that satisfies the requirements Item 6 in Lemma 5.D.1 is the sum of squares

$$L(\mathbf{t}) = \sum_m \sum_{i_1 \cdots i_m} |\mathbf{t}_m^{i_1 \cdots i_m}|^2$$

Which will be used to construct a loss function for tracks in Section 5.6.

5.E Details on experiments

The Brownian motion generated is always 2 dimensional and to a resolution of 100 time steps. All experiments were performed on a laptop CPU. For computing the Bayes' act we used stochastic gradient descent. In Sections 5.6.1 and 5.6.2 we use a low learning rate (.2) and many steps (50) to do this, but in Section 5.6.3 we use a high learning rate (5) and few steps (10) to speed things up as the accuracy is barely affected. Depending on the experiments, we computed the signature up to degree 6, but sometimes lower. The optimal bandwidth for the kernel was chosen from $10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}$.

Permutation testing The permutation testing was performed by permuting the data 50 times and checking if the unpermuted statistic lies within the 5:th percentile or not. This was done 50 times with new samples from the distributions to get an accuracy, and this procedure was repeated twice to get a range of values.

Discretization and Interpolation. As discussed in the main text, given $\mathbf{x}(t_1), \dots, \mathbf{x}(t_L) \in \mathbb{R}^d$ we identify them as elements of the path $t \mapsto \mathbf{x}(t)$ given by piecewise linear interpolation. Other choices are possible but ultimately the specific choice does not matter when the underlying path is smooth since they all converge to the same element of $\mathcal{H}$ where the iterated integrals are defined in the classical Riemann-Stieltjes sense. When the underlying path is not smooth, e.g. a Brownian motion, the situation is more subtle and in general the interpolation choice matters: for example, piecewise constant interpolations would lead to iterated Ito integrals and piecewise linear would result in iterated Stratonovich integrals. However, for any "reasonable interpolation" to a continuous path, the specific choice doesn't matter, see [74] for a detailed discussion. Among all such "reasonable interpolations", the piecewise linear one is arguably the most canonical if we do not have additional information about the behaviour in between time points.

Besides the choice of interpolation scheme one can also ask about the effects of the time discretization. A detailed study of this error would require model assumption and is beyond the scope of this article. However, to provide some intuition revisit the setup of Experiment 5.6.1 and plot the quantity

$$n \mapsto d(\delta_{\mathbf{x}_n}, \delta_{\mathbf{y}_n})$$

where $\mathbf{x}_n$ is the piecewise linear path given by interpolation of a Brownian trajectory $t \mapsto \mathbf{x}(t)$ along the points $t_i^n := \frac{i}{n}$ for $i = 0, \dots, n$; y_n is the same but for the

time-changed trajectory $t \mapsto \mathbf{y}(t) = \mathbf{x}(\varphi(t))$ with the time-change φ as detailed in Section 5.6. Figure 5.7 shows how the sampling error decreases as the discretization error vanishes.

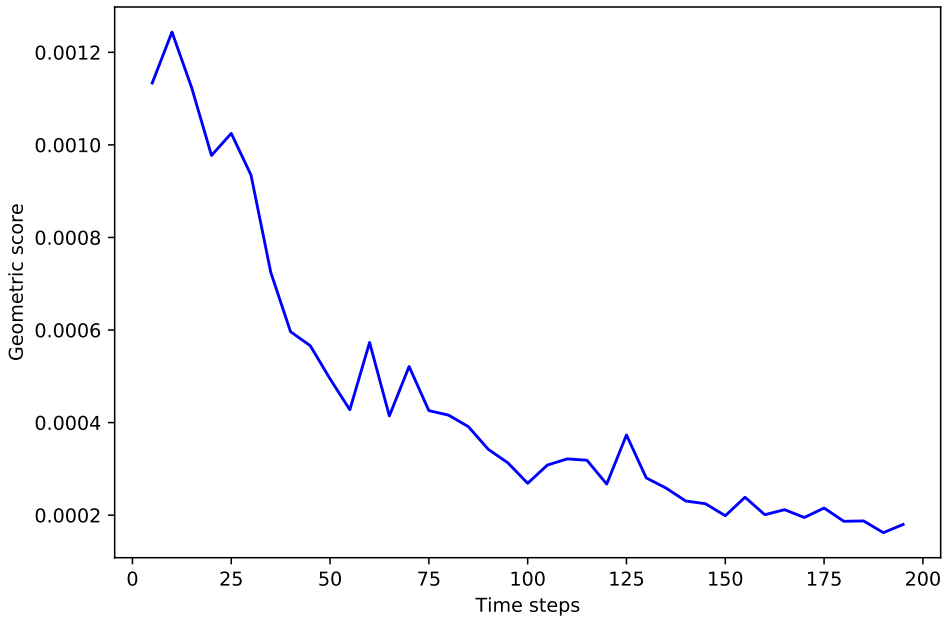


Figure 5.7: The geometric score of a Brownian path compared to a warped version of itself as the number of time steps n vary from 5 to 200. The plot shows the mean over 100 samples with time warp $\varphi(t) = t^{25}$.

5.F Pansu differentiability, convexity and gradient descent on Carnot groups

Recall that a Carnot group G comes equipped with a dilation operator

$$\delta_\lambda : G \to G, (x_0, x_1, \dots, x_n) \mapsto (x_0, \lambda x_1, \dots, \lambda^n x_n).$$

Definition 47. The Pansu derivative of a function $f : G \to \mathbb{R}$ is defined as (whenever it exists)

$$Df(g)h = \lim_{\lambda \downarrow 0} \frac{f(g\delta_\lambda h) - f(g)}{\lambda}$$

If f has a Pansu derivative for all $g \in G$ then we say that f is Pansu differentiable.

Note that $Df(g)$ takes values in $\mathbb{R}^d$, hence we may map it back to the group by exponentiating. The idea of this chapter was that by using update rules of the form

$$g_{i+1} = g_i e^{-\eta Df(g_i)}$$

one is able to update the function f directly on the group G independent of its vector space embedding and without projection methods. This was stated in Section 5.5 but was not explored in depth. The difficulty here was finding a notion of convexity on carnot groups that satisfies both of the following criteria

1. For a convex function the update rule converges to a unique minimum.

2. There are a lot of convex functions.

We explored three different notions of convexity: strong convexity, h-convexity, and dilative convexity. They all have different strengths and weaknesses but ultimately they all fail to satisfy both of the above requirements. Note that by Theorem 5.5.1, the update rule is always decreasing so the only thing we need to check for Item 1 is that a convex function has a unique minimum where its Pansu gradient is 0.

5.F.1 Strong Convexity

Definition 48. We say that a function $u : G \to \mathbb{R}$ is strongly convex if for any $g, h \in G$ and $0 \leq \lambda \leq 1$

$$u(g\delta_\lambda(g^{-1}h)) \leq u(g) + \lambda(u(h) - u(g)).$$

Proposition 5.F.1. *If $u : G \to \mathbb{R}$ is strongly convex then any local minimum is a global minimum.*

Proof. Let g, g' be two local minima such that $u(g) \leq u(g')$, then

$$u(g\delta_\lambda(g^{-1}g')) \leq u(g) + \lambda(u(g') - u(g))$$

hence the only way that g is a local minimum is if $u(g) = u(g')$ and $u(g\delta_\lambda(g^{-1}g'))$ lies in the same level set for every $0 \leq \lambda \leq 1$. □

Example 5.F.1. Let $u : G \to \mathbb{R}$ be a function of only the first level of G, then u is strongly convex if and only if it is linearly convex. This holds since on the first level

$$u(g\delta_\lambda(g^{-1}g')) = u(g + \lambda(g' - g)).$$

Unfortunately we were unable to find any non-trivial examples of strongly convex functions besides the above one.

5.F.2 h-convexity

If one is slightly more restrictive in defining the derivative, then the corresponding convexity class is much larger.

Definition 49. We say that a function $u : G \to \mathbb{R}$ is h-convex if for any $g \in G$, $v \in \mathbb{R}^d$ and $0 \leq \lambda \leq 1$

$$u(ge^{\lambda v}) \leq u(g) + \lambda(u(ge^v) - u(g)).$$

Remark 5.F.1. This is equivalent to requiring that the definition for strong convexity holds for h of the form $h = ge^v$

Remark 5.F.2. This is also equivalent to requiring that the function $f(t) = u(ge^{\lambda tv})$ is convex for every $v \in \mathbb{R}^d$.

Definition 50. For a function $u : G \to \mathbb{R}$ we define its geometric or pansu gradient $\nabla u(g)$ as

$$\langle \nabla u(g), v \rangle = \lim_{\varepsilon \to 0} \frac{u(ge^{\varepsilon v}) - u(g)}{\varepsilon}$$

The geometric Hessian Δu is defined by $\Delta u = \nabla \circ \nabla u$ and the symmetric version $\Delta_S u$ is defined by $\Delta_S u = Sym(\Delta u)$

Theorem 5.F.1 (Theorem 5.12 [55])**.** *if $u \in C^2$, then u is h-convex iff $\Delta_S u$ is positive semidefinite*

Lemma 5.F.2. *1. ∇ is linear, chain rule and product rule applies.*

2. *if $u(g) = \langle v_1 \otimes \cdots \otimes v_n, g \rangle$ then $\nabla u(g) = \langle v_1 \otimes \cdots \otimes v_{n-1}, g \rangle v_n$*

3. *if $h : \mathbb{R}^n \to \mathbb{R}$ is convex and non-decreasing and $u : G \to \mathbb{R}^n$ is h-convex then $h \circ u$ is h-convex*

4. *the h-derivative can be represented in coordinates as*

$$\partial_{x_i} = \sum_{i_1 \cdots i_k} g^{i_1 \cdots i_k} \partial_{i_1 \cdots i_k i}$$

Example 5.F.2. If $U(g) = u(x, y, t)$ is a function on a 2-step group of dimension 2, then

$$\nabla U = \begin{pmatrix} \partial_x u - y \partial_t u \\ \partial_y u + x \partial_t u \end{pmatrix}$$

so if $u(x,y,t) = \left((x^2+y^2)^2+16t^2\right)^{1/4}$ is the gauge function of the step 2 group, then u is h-convex by [55, Theorem 6.6], and its horisontal gradient is given by

$$\nabla U = U(g)^{-3}\left((x^2+y^2)\begin{pmatrix}x\\y\end{pmatrix} + 8t\begin{pmatrix}-y\\x\end{pmatrix}\right).$$

If we fix some g_0 and let $\Delta_x = x - x_0, \Delta_y = y - y_0$ and consider $U(g_0^{-1}g)$ then

$$\nabla U = U(g)^{-3}\left((\Delta_x^2+\Delta_y^2)\begin{pmatrix}\Delta_x\\\Delta_y\end{pmatrix} + 8(t-t_0+y_0x-x_0y)\begin{pmatrix}-\Delta_y\\\Delta_x\end{pmatrix}\right)$$

h-convexity of linear systems

Lemma 5.F.3. *The function $t \mapsto \|e^{tA}v\|^2$ is convex for every $v \in \mathbb{R}^d$ iff $A^TA + A^2 \succcurlyeq 0$*

Proof. If $f(t) = \|e^{tA}v\|^2$ then $\dot{f}(t) = 2\langle e^{tA}v, Ae^{tA}v\rangle$ and $\ddot{f}(t) = 2\|Ae^{tA}v\|^2 + 2\langle e^{tA}v, A^2e^{tA}v\rangle$. The reverse implication follows since e^{tA} is invertible. □

Definition 51. We say that a matrix A is *quasi-definite* (q.d) if $A^2 + A^TA \succcurlyeq 0$, or equivalently if $(A+A^T)^2 + A^TA - AA^T$ has only non-negative eigenvalues.

Proposition 5.F.4. *Let $A : \mathbb{R}^m \otimes \mathbb{R}^d \to \mathbb{R}^d$ be a linear map such that Aw is q.d for every $w \in \mathbb{R}^m$. Then the ODE map $F(x) = \|y_T\|^2$ where $y(0) = y_0$ and $dy_s = A(y_s \otimes dx_s)$ is h-convex for any $y_0 \in \mathbb{R}^d$.*

Proof. Follows from the lemma since $F(x \star v_t) = \|e^{Avt}y_T\|^2$. □

Remark 5.F.3. If the map $F(x) = \|y_T\|^{2p}, p > 1$ is used instead of the squared norm then the condition on A is unchanged since if $f(t) = \|e^{tA}v\|^{2p}$ then $\dot{f}(t) = 2p\langle e^{tA}v, e^{tA}v\rangle^{p-1}\langle e^{tA}v, Ae^{tA}v\rangle$ and

$$\begin{aligned}\ddot{f}(t) &= 4p(p-1)\langle e^{tA}v, e^{tA}v\rangle^{p-2}\langle e^{tA}v, Ae^{tA}v\rangle^2 + 2p\langle e^{tA}v, e^{tA}v\rangle^{p-1}\langle e^{tA}v, (A^TA+A^2)e^{tA}v\rangle \\ &= 2p\|e^{tA}v\|^{2p-4}\Big(2(p-1)\langle e^{tA}v, Ae^{tA}v\rangle^2 + \|e^{tA}v\|^2\langle e^{tA}v, (A^TA+A^2)e^{tA}v\rangle\Big)\end{aligned}$$

This looks like a more relaxed condition and one might be tempted into thinking that this becomes more convex as p increases. This is not true however since for $A = \left(\begin{smallmatrix}1 & 1\\ -1 & 1\end{smallmatrix}\right)$ it holds that

$$\begin{aligned}&2(p-1)\langle e^{tA}v, Ae^{tA}v\rangle^2 + \|e^{tA}v\|^2\langle e^{tA}v, (A^TA+A^2)e^{tA}v\rangle \\ &= 2(p-1)v_1^4 + 2(v_1^2+v_2^2)v_1^2 + 4(v_1^2+v_2^2)v_1v_2\end{aligned}$$

which is indefinite for every p.

Remark 5.F.4. Clearly if $A^2 \succcurlyeq 0$ then A is q.d. If A is normal then clearly $(A+A^T)^2 + A^TA - AA^T = (A+A^T)^2 \succcurlyeq 0$ so A is q.d, note that $A^TA - AA^T$ is always (in dim 2) indefinite

Example 5.F.3 (Rotation scaling)**.** Let $A = \begin{pmatrix} r & -1 \\ 1 & r \end{pmatrix}$ so that

$$A^2 = \begin{pmatrix} r^2-1 & -2r \\ 2r & r^2-1 \end{pmatrix}, \quad A^TA = \begin{pmatrix} r^2+1 & 0 \\ 0 & r^2+1 \end{pmatrix}$$

so that A is q.d for every $r \in \mathbb{R}$, but $A^2 \succcurlyeq 0$ only when $|r| \geq 1$.

Proposition 5.F.5. *For every A and diagonal D with full rank there exists some $\eta > 0$ such that $A + \eta D$ is q.d.*

Proof. Assume WLOG that $D = I$ is the identity matrix. If $\|Av\| \leq L\|v\|$, then $|v^TAv| \leq L\|v\|^2$. Note that if $B = A + \eta D$, then $B^2 = A^2 + 2\eta AD + \eta^2 D^2$. So for any $v \in \mathbb{R}^d$ we can write

$$v^TB^2v = \eta^2\|v\|^2 + 2\eta v^TAv + v^TA^2v \geq (\eta^2 - 2L\eta - L^2)\|v\|^2.$$

Hence any $\eta \geq L$ works. □

Proposition 5.F.6 (Regularisation)**.** *If $A \in L(\mathbb{R}^d \otimes \mathbb{R}^m, \mathbb{R}^d)$, Let $F(x)$ be the ODE map*

$$F(x) = \|y_T\|^2, \quad dy = Aydx, \quad y(0) = y_0,$$

and for $\eta > 0$ denote by $F_\eta(x)$ the nonlinear ODE map

$$F_\eta(x) = \|y_T\|^2, \quad dy = (Ay + \eta y\frac{\dot{x}^T}{\|\dot{x}\|})dx, \quad y(0) = y_0,$$

Then:

1. $F_\eta(x) = e^{2\eta\int\|\dot{x}\|dt}F(x)$
2. *For η large enough, F_η is h-convex.*

Proof. Note that if y is the solution for a piecewise linear $x = (\Delta_1, \ldots, \Delta_n)$ and map A we can write

$$y_T = e^{A\Delta_n} \cdots e^{A\Delta_1} y_0$$

and if $\tilde{y}$ is the regularised solution we can write

$$\begin{aligned}\tilde{y}_T &= e^{A\Delta_n+\eta\frac{\langle\Delta_n,\dot{x}_n\rangle}{\|\dot{x}_n\|}I}\cdots e^{A\Delta_1+\eta\frac{\langle\Delta_n,\dot{x}_n\rangle I}{\|\dot{x}_n\|}I}y_0\\ &= e^{\eta\|\dot{x}_n\|dt_n+\cdots+\eta\|\dot{x}_1\|dt_1}e^{A\Delta_n}\cdots e^{A\Delta_1}y_0 = e^{\eta\int\|\dot{x}\|^2dt}y_T.\end{aligned}$$

By continuity this holds for all C^1 paths x. To see that F_η is h-convex we need to show that we can choose η large enough that $(Aw+\eta\|w\|I)^2 \succcurlyeq 0$ uniformly in $w\in\mathbb{R}^m$. Just like in the previous proposition we can pick $v\in\mathbb{R}^d$ and write

$$v^T(Aw+\eta\|w\|I)^2v = \eta^2\|w\|^2\|v\|^2 + 2\eta\|w\|v^T(Aw)v + v^T(Aw)^2v \geq (\eta^2-2\eta L-L^2)\|w\|^2\|v\|^2$$

where L is the Lipschits constant of A, and $\eta\geq L$ works. □

Corollary 5.F.7. *If $x\sim z$, iff $F_\eta(xz^{-1})=0$ for every A.*

Remark 5.F.5. Since F_η is nonlinear in x we could have $x\sim z$ but $F_\eta(x)\neq F_\eta(z)$.

Remark 5.F.6. The regularisation cannot be done linearly in x! If we regularise with a linear function ℓ in x then we run into issues where the linear function is not bounded from below in the subspace spanned by $\ell^\perp$ so convexity can not be guaranteed. Note also that it's important that the regularisation maps into multiples of the identity map since these are the only ones that commute with all other matrices. If we used a family of linear functions to embed $\mathbb{R}^m$ into $\mathbb{R}^d$ (assuming $d\geq m$). This would probably yield a $q.d$ map that is TLE, however there would be no simple guarantee that these characterise the paths.

No uniqueness guarantee

Unfortunately h-convexity proves inadequate In the very simple way that it does not guarantee a unique minimum. The definition only requires convexity in horisontal directions which makes it easy to check and qualify, but it means that a proof like in Proposition 5.F.1 becomes impossible.

5.F.3 Dilative convexity

There is another notion of convex combinations on the group, and this leads to a different class of convex functions.

Definition 52. The convex combination of g and h is defined as

$$c_\lambda(g,h) := \delta_{1-\lambda}g\delta_\lambda h$$

This new type of convex combination corresponds to scaling down g and adding a scaled down version of h, whereas the one used in previous sections was to move from g to a scaled down version of $g^{-1}h$. Unlike the previous convex combination, this one has the symmetric property $c_\lambda(g,h)^{-1} = c_{1-\lambda}(g^{-1},h^{-1})$

Definition 53. We say that a function $f : G \to \mathbb{R}$ is convex if $f(c_\lambda(g,h)) \leq f(g) + \lambda(f(h) - f(g)$

Proposition 5.F.8. *1. Any subadditive homogenous function is convex*

2. if f is convex, its level sets are convex

3. if f is strictly convex, any local minimum is a global minimum

Proof. Note that if f is subadditive and homogenous, then

$$f(\delta_{1-\lambda} g \delta_\lambda h) \leq f(\delta_{1-\lambda}) + f(\delta_\lambda h) = (1-\lambda)f(g) + \lambda f(h).$$

2 and 3 are clear. □

On the surface this looks good, there are a lot of functions that satisfy this convexity requirement, and they all have unique minima. The problem with this class is slightly more subtle in that there is no guarantee that having 0 gradient implies that you are at the minimum. The two previous classes enjoyed the fact that the convex combinations converge to the Pansu derivative for small values of λ. This notion of convex combination unfortunately does not have this.

Definition 54. Recall that if f is C^1 on G we define its pansu derivative as

$$\nabla_P^h f(g) = \lim_{\varepsilon \to 0} \frac{f(g\delta_\varepsilon h) - f(g)}{\varepsilon}.$$

Additionally, we define the twisted gradient as

$$\nabla_T^h f(g) = \lim_{\varepsilon \to 0} \frac{f(\delta_{1-\varepsilon} g \delta_\varepsilon h) - f(g)}{\varepsilon}.$$

Proposition 5.F.9. *1. For coordinates $I = i_1 \cdots i_n$ and a coordinate mapping e_I it holds that*

$$\nabla_P^h e_I(g) = e_{i_1 \cdots i_{n-1}}(g) e_{i_n}(h),$$
$$\nabla_T^h e_I(g) = \nabla_P^h e_I(g) - n e_I(g)$$

2. *If* $F : \mathbb{R}^k \to \mathbb{R}$ *and* $f(g) = F(e_{I_1}(g), \dots, e_{I_k}(g))$ *is a function of coordinates, then*

$$\nabla_P^h f(g) = \sum_{i=0}^{k} \partial_i F \nabla_P^h e_{I_i}(g),$$
$$\nabla_T^h f(g) = \sum_{i=0}^{k} \partial_i F \nabla_T^h e_{I_i}(g)$$

3. *For any* C^1 *function* f *it holds that* $\nabla_P f(0) = \nabla_T f(0)$.

4. *A* C^1 *function* f *is convex if and only if for every* $g, h \in G$

$$f(g) + \nabla_T^h f(g) \leq f(h)$$

5. *Define for* $t \in T$, $(St)^n = nt^n$, *then*

$$\nabla_T^h f(g) = \nabla_P^h f(g) - \nabla f(g) \cdot Sg.$$

Also if we define the linear map $\delta g : G^\star \to \mathbb{R}^d$ *by* $(\delta g)_{Ii,i} = g^I$ *then we can further write*

$$\nabla_T^h f(g) = \nabla f(g)(\delta g \cdot h - Sg).$$

6. *If* f *is homogeneous and* C^1, *then* $\nabla_p^h f(g) = \nabla_P^{\delta_\lambda h} f(\delta_\lambda g)$

0 gradient does not guarantee a minimum

Remark 5.F.7. Unlike the pansu gradient, the twisted gradient is not a linear map, but an affine one, and no statement of the following type holds

$$f(g) + \nabla_P^h f(g) \leq f(h)$$

but instead you get

$$f(g) + \nabla_T^h f(g) \leq f(h).$$

Example 5.F.4. The squared regular homogenous norm $f(g) = \sum_n \|g_n\|^{2/n}$ satisfies

$$\nabla_T^h f(g) = \nabla_P^h f(g) - 2f(g)$$

so that the above inequality turns into

$$\nabla_P^h f(g) \leq f(h) + f(g)$$

and if $\nabla_P^h f(g) = 0$ this only tells us that $f(h) + f(g) \geq 0$.

Because of this, having 0 Pansu gradient doesn't guarantee that you are at a minimum, so the update rule is not guaranteed to converge to a minimum.

Closing thoughts

To summarise, the main contributions of this thesis have been new algebraic and combinatorial methods of studying and characterising measures, particularly measures on paths. The signature cumulants from Chapter 1 ([30]) introduces a new way to study joint measures on paths and their independence by introducing the rich combinatorial structure of ordered partitions. The kernelised cumulants from Chapter 2 ([31]) extends classical cumulants to reproducing kernel Hilbert spaces, making them tractable in high dimensions, and allowing them to benefit from techniques from kernel learning. This is achieved by introducing partition measures, which allow one to elegantly express these cumulants as sums of product kernels. Chapter 3 ([189]) generalises discrete-time signatures to a wider range of signature-esque maps by extending algebraic embeddings from state spaces to embeddings on time series on that space. These embeddings are just as powerful as the signature, but can be computed much more cheaply. In Chapter 4 ([28]) we use the structure of higher rank tensor algebras to study higher rank signatures which induce topologies that are richer than the weak topology and are able to characterise measures on paths in a much stronger sense. The inducing map is also useful as a feature map allowing for regression and similar techniques. Chapter 5 ([29]) introduces a new family of scoring rules for paths up to reparametrisation by relying on their group structure. This allows one to study divergences, entropy and mutual information up to reparametrisation in a structured way.

Open questions:

- There is a general correspondence between different types of cumulants and sets of partitions, see for instance [156, Section 6], but the signature cumulants and ordered partitions do not fit neatly into this framework and it would be interesting to understand if there is a way to unify the two approaches.
- The higher rank signatures were studied in discrete time but could in theory be extended to continuous time. Many of the definitions and results extend to

continuous time via rough paths where the sums in the discrete higher rank signatures turn into iterated (rough) integrals and even the equivalence of items 1, 2, and 3 in Theorem 4.4.2 follows with minor modifications of the given proof. However, the main hurdle in going to continuous time is encountered even for $r = 1$: the prediction process $t \mapsto \hat{\mathbf{X}}^1_t = \mathbb{P}(X \in \cdot | \mathcal{F}_t)$ generally only has càdlàg trajectories, even if the sample paths of $t \mapsto X_t$ are continuous. Without any additional assumptions, this naturally requires structures that appear in càdlàg rough path theory which is still an area of ongoing research; see for example work of Chevyrev, Friz, and Shekhar [42, 77].

- Also for the higher rank signatures, there are currently no good bounds for how much one gains by going higher up in the order of topologies, or how to truncate the higher order signatures efficiently. Similarly, there are no rates of convergence when estimating the metrics from samples, unlike the case of adapted Wasserstein distances where these are well known, see [7].

- By using the results from [113], and combining it with the results from Chapter 2 one may construct a set of cumulants on signatures of paths as the kernelised cumulants with a signature kernel. These will be different from, but have similar properties to the signature cumulants and it would be interesting to know what the advantages/disadvantages are of using either of the two.

- Most of the power of the results in Chapter 3 comes from the fact that it is more efficient to compute compositions of many simple transformations rather than one complicated one, but this is not quantified and it would be interesting to understand this better even in simple cases such as computing polynomials.

- In Chapter 5 we explore a gradient descent update rule where one traverses the group directly by multiplicative steps, but we were unable to find a natural class of functions that behave well with respect to this update rule. This is explained more in Section 5.F, but such a class would need to be structured enough that its members converge with this update rule, but big enough to contain common loss functions for optimisation.

Bibliography

[1] J. Abernethy and Rafael M. Frongillo. A characterization of scoring rules for linear properties. In *COLT*, 2012.

[2] Marcelo Aguiar, Nantel Bergeron, and Frank Sottile. Combinatorial hopf algebras and generalized dehn–sommerville relations. *Compositio Mathematica*, 142(01):1–30, January 2006.

[3] D. J. Aldous. Weak convergence and general theory of processes. *Unpublished draft of monograph*, 1981.

[4] Miguel A. Arcones and Eva Giné. On the bootstrap of U and V statistics. *Annals of Statistics*, 20:655–674, 1992.

[5] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[6] J. Backhoff-Veraguas, D. Bartl, M Beiglböck, and M Eder. All adapted topologies are equal. *arXiv preprint arXiv:1905.00368*, 2019.

[7] J. Backhoff-Veraguas, D. Bartl, M Beiglböck, and J. Wiesel. Estimating processes in adapted wasserstein distance. *arXiv preprint arXiv:2002.07261*, 2020.

[8] Pierre Baldi, Søren Brunak, Paolo Frasconi, Giovanni Soda, and Gianluca Pollastri. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, 15(11):937–946, 1999.

[9] Ludwig Baringhaus and C. Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88:190–206, 2004.

[10] D. Bartl, M. Beiglboeck, and G. Pammer. The wasserstein space of filtered processes. *arXiv:2104.14245*, 2021.

[11] Mustafa Baydogan. Multivariate time series classification datasets. http://mustafabaydogan.com, 2015. [Accessed: 2020-06-11].

[12] Mustafa Gokce Baydogan and George Runger. Learning a symbolic representation for multivariate time series classification. *Data Mining and Knowledge Discovery*, 29(2):400–422, 2015.

[13] Mustafa Gokce Baydogan and George C. Runger. Time series representation and similarity based on local autopatterns. *Data Mining and Knowledge Discovery*, 30:476–509, 2015.

[14] Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *The Journal of Machine Learning Research*, 18(1):2653–2688, 2017.

[15] Alessio Benavoli, Giorgio Corani, and Francesca Mangili. Should we really use post-hoc tests based on mean-ranks? *The Journal of Machine Learning Research*, 17(1):152–161, January 2016.

[16] Alessio Benavoli, Giorgio Corani, Francesca Mangili, Marco Zaffalon, and Fabrizio Ruggeri. A bayesian wilcoxon signed-rank test based on the dirichlet process. In *International conference on machine learning*, pages 1026–1034, 2014.

[17] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

[18] D. P. Bertsekas and S. E. Shreve. *Stochastic optimal control. The discrete time case*. Academic Press. New York, 1978.

[19] Hans Bühler, Blanka Horvath, Terry Lyons, Imanol Perez Arribas, and Ben Wood. A data-driven market simulator for small data environments, 2020.

[20] Patrick Billingsley. *Probability and Measure*. John Wiley and Sons, 1986.

[21] J Bion-Nadal and D Talay. On a wasserstein-type distance between solutions to stochastic differential equations. *Annals of Applied Probability*, 2019.

[22] Mathieu Blondel, A. Mensch, and Jean-Philippe Vert. Differentiable divergences between time series. In *AISTATS*, 2021.

[23] H. Boedihardjo, X. Geng, T. J. Lyons, and D. Yang. The signature of a rough path: Uniqueness. *Advances in Mathematics*, 2016.

[24] V. I Bogachev. *Measure theory. Vol. I, II.* Springer, 2007.

[25] David Bolin and Finn Lindgren. Excursion and contour uncertainty regions for latent gaussian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1):85–106, 2015.

[26] Silvere Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, September 2013.

[27] Patric Bonnier, Patrick Kidger, Imanol Perez Arribas, Cristopher Salvi, and Terry Lyons. Deep signature transforms. *33rd Conference on Neural Information Processing Systems, NeurIPS*, 2019.

[28] Patric Bonnier, Chong Liu, and Harald Oberhauser. Adapted topologies and higher rank signatures. *In press, Annals of Applied Probability*, 2021.

[29] Patric Bonnier and Harald Oberhauser. Proper scoring rules, gradients, divergences, and entropies for paths and time series. *arXiv preprint arXiv:2111.06314*.

[30] Patric Bonnier and Harald Oberhauser. Signature cumulants, ordered partitions, and independence of stochastic processes. *Bernoulli*, 26(4), November 2020.

[31] Patric Bonnier, Harald Oberhauser, and Zoltan Szábó. Kernelized cumulants: Beyond kernel mean embeddings. *arXiv preprint arXiv:2301.12466*, 2023.

[32] Glenn W Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[33] Graham Brightwell and Peter Winkler. Counting linear extensions. *Order*, 8(3):225–242, 1991.

[34] Roger W. Brockett. Volterra series and geometric control theory. *Automatica*, 12(2):167 – 176, 1976.

[35] D.S. Broomhead and D. Lowe. Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2:321–355, 1988.

[36] J Douglas Carroll and Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-young" decomposition. *Psychometrika*, 35(3):283–319, 1970.

[37] Thomas Cass and Peter Friz. Malliavin calculus and rough paths. *Bulletin des Sciences Mathématiques*, 135(6):542–556, 2011. Special issue in memory of Paul Malliavin.

[38] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

[39] K. T. Chen. Iterated integrals and exponential homomorphisms. *Proc. London Math. Soc, 4, 502–512*, 1954.

[40] K. T. Chen. Integration of paths, geometric invariants and a generalized Baker-Hausdorff formula. *Ann. of Math. (2), 65:163–178*, 1957.

[41] K. T Chen. Integration of paths – a faithful representation of paths by non-commutative formal power series. *Transactions of the American Mathematical Society*, 1958.

[42] I Chevyrev and P Friz. Canonical rdes and general semimartingales as rough paths. *Annals of probability*, 2019.

[43] I. Chevyrev and A. Kormilitzin. A primer on the signature method in machine learning. *arXiv preprint arXiv:1603.03788*, 2016.

[44] I Chevyrev and T. J Lyons. Characteristic functions of measures on geometric rough paths. *Annals of Probability*, 2016.

[45] 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, Jan 2020.

[46] Ilya Chevyrev and Harald Oberhauser. Signature moments to characterize laws of stochastic processes. *arXiv e-prints*, page arXiv:1810.10971, Oct 2018.

[47] Wei-Liang Chow. Über Systeme von linaren partiellen Differentialgleichungen erster Ordnung. *Mathematische Annalen*, 117-117(1):98–105, December 1940.

[48] EunYi Chung and Joseph P. Romano. Exact and asymptotically robust permutation tests. *Annals of Statistics*, 41(2):484–507, 2013.

[49] Andrzej Cichocki, Namgil Lee, Ivan Oseledets, Anh-Huy Phan, Qibin Zhao, and Danilo P Mandic. Tensor networks for dimensionality reduction and large-scale optimization: Part 1 low-rank tensor decompositions. *Foundations and Trends® in Machine Learning*, 9(4-5):249–429, 2016.

[50] Nadav Cohen, Or Sharir, and Amnon Shashua. On the expressive power of deep learning: A tensor analysis. In *Conference on learning theory*, pages 698–728, 2016.

[51] Chris Cremer, Xuechen Li, and David Duvenaud. Inference suboptimality in variational autoencoders. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1078–1086, 2018.

[52] N Cristianini and J Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge, 2000.

[53] Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. In *ICML*, 2017.

[54] Marco Cuturi and Arnaud Doucet. Autoregressive Kernels For Time Series. *arXiv e-prints*, page arXiv:1101.0673, Jan 2011.

[55] Donatella Danielli, Nicola Garofalo, and Duy-Minh Nhieu. Notions of convexity in carnot groups. *Communications in Analysis and Geometry*, 11, 06 2003.

[56] A. Dawid and M. Musio. Theory and applications of proper scoring rules. *METRON*, 72:169–183, 2014.

[57] Philip Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59(1):77–93, 2007.

[58] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

[59] Joscha Diehl, Kurusch Ebrahimi-Fard, and Nikolas Tapia. Generalized iterated-sums signatures, 2020.

[60] Joscha Diehl, Kurusch Ebrahimi-Fard, and Nikolas Tapia. Time-warping invariants of multidimensional time series. *Acta Applicandae Mathematicae*, 170(1):265–290, May 2020.

[61] Richard Dudley. *Real Analysis and Probability*. Cambridge University Press, 2004.

[62] Joel Dyer, Patrick Cannon, and Sebastian M Schmon. Approximate bayesian computation with path signatures, 2021.

[63] Sathishkumar V E, Jangwoo Park, and Yongyun Cho. Using data mining techniques for bike sharing demand prediction in metropolitan city. *Computer Communications*, 153:353–366, 2020.

[64] K. Ebrahimi-Fard and F. Patras. Cumulants, free cumulants and half-shuffles. *Proceedings of the Royal Society*, 2015.

[65] Kurusch Ebrahimi-Fard, Frédéric Patras, Nikolas Tapia, and Lorenzo Zambotti. Hopf-algebraic Deformations of Products and Wick Polynomials. *International Mathematics Research Notices*, 12 2018.

[66] Manu Eder. Compactness in adapted weak topologies, 2019.

[67] Adeline Fermanian, Pierre Marion, Jean-Philippe Vert, and Gérard Biau. Framing rnn as a kernel method: A neural ode approach, 2021.

[68] Ricardo Pinto Ferreira. Combination of artificial intelligence techniques for prediction the behavior of urban vehicular traffic in the city of São Paulo. In *Anais do 10. Congresso Brasileiro de Inteligência Computacional*, pages 1–5, 2016.

[69] Tobias Fissler, Rafael Frongillo, Jana Hlavinová, and Birgit Rudloff. Forecast evaluation of quantiles, prediction intervals, and other set-valued functionals. *Electronic Journal of Statistics*, 15(1):1034–1084, 2021.

[70] Tobias Fissler and J. Ziegel. Order-sensitivity and equivariance of scoring functions. 2017.

[71] M. Fliess. Fonctionnelles causales non linéaires et indéterminées non commutatives. *Bull. Soc. Math. France*, 109(1):3–40, 1981.

[72] Michel Fliess. Un outil algebrique: Les series formelles non commutatives. In Giovanni Marchesini and Sanjoy Kumar Mitter, editors, *Mathematical Systems Theory*, pages 122–148, Berlin, Heidelberg, 1976. Springer Berlin Heidelberg.

[73] Vincent Fortuin, Dmitry Baranchuk, Gunnar Rätsch, and Stephan Mandt. GP-VAE: Deep probabilistic time series imputation. In *International Conference on Artificial Intelligence and Statistics*, pages 1651–1661. PMLR, 2020.

[74] Peter Friz and Harald Oberhauser. Rough path limits of the Wong-Zakai type with a modified drift term. *Journal of Functional Analysis*, 256(10):3236 – 3256, 2009.

[75] Peter K Friz, Paul Hager, and Nikolas Tapia. Unified signature cumulants and generalized magnus expansions. *arXiv preprint arXiv:2102.03345*, 2021.

[76] Peter K. Friz and M Hairer. *A Course on Rough Paths with an introduction to regularity structures*. Springer, 2014.

[77] Peter K. Friz and Atul Shekhar. General rough integration, lévy rough paths and a lévy-kintchine-type formula. *Ann. Probab.*, 45(4):2707–2765, 07 2017.

[78] Peter K. Friz and Nicolas B. Victoir. *Multidimensional stochastic processes as rough paths: theory and applications*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 2010.

[79] Rafael M. Frongillo and Ian A. Kash. Vector-valued property elicitation. In *COLT*, 2015.

[80] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems (NIPS)*, pages 498–496, 2008.

[81] Thomas Gärtner. A survey of kernels for structured data. *ACM SIGKDD Explorations Newsletter*, 5(1):49–58, 2003.

[82] R Giles. A generalization of the strict topology. *Transactions of the American Mathematical Society*, 1971.

[83] Chad Giusti and Darrick Lee. Signatures, lipschitz-free spaces, and paths of persistence diagrams, 2021.

[84] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[85] AL Goldberger, LAN Amaral, L Glass, JM Hausdorff, P Ch Ivanov, RG Mark, JE Mietus, GB Moody, CK Peng, and HE Stanley. Components of a new research resource for complex physiologic signals. *PhysioBank, PhysioToolkit, and Physionet*, 2000.

[86] I. J. Good. Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 14(1):107–114, 1952.

[87] Benjamin Graham. Sparse arrays of signatures for online character recognition. *arXiv preprint arXiv:1308.0371*, 2013.

[88] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE workshop on automatic speech recognition and understanding*, pages 273–278. IEEE, 2013.

[89] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.

[90] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602 – 610, 2005.

[91] F. Gressmann, F. J. Király, B. Mateen, and H. Oberhauser. Probabilistic supervised learning. *ArXiv e-prints*, January 2018.

[92] 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.

[93] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory (ALT)*, pages 63–78, 2005.

[94] Arthur Gretton, Kenji Fukumizu, Choon H Teo, Le Song, Bernhard Schölkopf, and Alex J Smola. A kernel statistical test of independence. In *Advances in neural information processing systems*, pages 585–592, 2008.

[95] Peter D. Grunwald and A. Dawid. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *Annals of Statistics*, 32:1367–1433, 2004.

[96] Peter Hall and Nader Tajvidi. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):359–374, 2002.

[97] B. M Hambly and T. J Lyons. Uniqueness for the signature of a path of bounded variation and the reduced path group. *Annals of Mathematics*, 2010.

[98] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[99] Marti A. Hearst. Support vector machines. *IEEE Intelligent Systems*, 13(4):18–28, July 1998.

[100] M. F. Hellwig. Sequential decisions under uncertainty and the maximum theorem. *Journal of Mathematical Economy*, 1996.

[101] D. Hoover and J. Keisler. Adapted probability distributions. *Transactions of the American Mathematical Society*, 1984.

[102] Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. In *ICML*, 2020.

[103] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

[104] Ming Hou, Jiajia Tang, Jianhai Zhang, Wanzeng Kong, and Qibin Zhao. Deep multimodal multilinear fusion with high-order polynomial pooling. In *Advances in Neural Information Processing Systems*, pages 12136–12145, 2019.

[105] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963, Jul 2019.

[106] G Jenča and P Sarkoci. Linear extensions and order-preserving poset partitions. *Journal of Combinatorial Theory, Series A*, 2014.

[107] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Samuel Harford. Multivariate lstm-fcns for time series classification. *Neural Networks*, 116:237 – 245, 2019.

[108] Isak Karlsson, Panagiotis Papapetrou, and Henrik Boström. Generalized random shapelet forests. *Data Min. Knowl. Discov.*, 30(5):1053–1085, September 2016.

[109] Valentin Khrulkov, Oleksii Hrinchuk, and Ivan Oseledets. Generalized tensor models for recurrent neural networks. *arXiv preprint arXiv:1901.10801*, 2019.

[110] Patrick Kidger, James Foster, Xuechen Li, Harald Oberhauser, and Terry Lyons. Neural sdes as infinite-dimensional gans, 2021.

[111] Patrick Kidger and Terry Lyons. Signatory: differentiable computations of the signature and logsignature transforms, on both CPU and GPU. In *International Conference on Learning Representations*, 2021. https://github.com/patrick-kidger/signatory.

[112] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[113] Franz J Király and Harald Oberhauser. Kernels for sequentially ordered data. *Journal of Machine Learning Research*, 2019.

[114] Lev Klebanov. *N-Distances and Their Applications*. Charles University, Prague, 2005.

[115] Jean Kossaifi, Zachary C Lipton, Aran Khanna, Tommaso Furlanello, and Anima Anandkumar. Tensor regression networks. *arXiv preprint arXiv:1707.08308*, 2017.

[116] Nils M. Kriege, Fredrik D. Johansson, and Christopher Morris. A survey on graph kernels. *Applied Network Science*, 5(1), 2020.

[117] Serge Lang. *Algebra*, volume 211. Springer Science & Business Media, 2012.

[118] R Lasalle. Causal transference plans and their monge-kantorovich problems. *Stochastic Analysis and Applications*, 2018.

[119] Darrick Lee and Robert Ghrist. Path signatures on lie groups, 2020.

[120] F. Lehner and T. Hasebe. Cumulants, spreadability and the campbell-baker-hausdorff series. *arXiv preprint arXiv:1711.00219*, 2017.

[121] Maud Lemercier, Cristopher Salvi, Thomas Cass, Edwin V. Bonilla, Theodoros Damoulas, and Terry Lyons. Siggpde: Scaling sparse gaussian processes on sequential data, 2021.

[122] C Leslie and R Kuang. Fast string kernels using inexact matching for protein sequences. *Journal of Machine Learning Research*, 2004.

[123] Yingzhen Li and Stephan Mandt. Disentangled sequential autoencoder, 2018.

[124] Paul Pu Liang, Zhun Liu, Yao-Hung Hubert Tsai, Qibin Zhao, Ruslan Salakhutdinov, and Louis-Philippe Morency. Learning representations from imperfect time series data via tensor rank regularization. *arXiv preprint arXiv:1907.01011*, 2019.

[125] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9628–9639, Red Hook, NY, USA, 2018. Curran Associates Inc.

[126] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*, 2018.

[127] Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. Text classification using string kernels. *The Journal of Machine Learning Research*, 2:419–444, 2002.

[128] Russell Lyons. Distance covariance in metric spaces. *The Annals of Probability*, 41:3284–3305, 2013.

[129] T. J Lyons, M Caruana, and T Lévy. *Differential equations driven by rough paths*. Springer, 2007.

[130] Terry Lyons. Rough paths, signatures and the modelling of functions on streams. *arXiv preprint arXiv:1405.4537*, 2014.

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

[132] Natsumi Makigusa. Two-sample test based on maximum variance discrepancy. Technical report, 2020. (https://arxiv.org/abs/2012.00980).

[133] Peter McCullagh. *Tensor Methods in Statistics*. Courier Dover Publications, 2018.

[134] Ilya Molchanov and Ilya S Molchanov. *Theory of random sets*, volume 87. Springer, 2005.

[135] Richard Montgomery. *A tour of subriemannian geometries, their geodesics and applications*, volume 91 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2002.

[136] 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.

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

[138] Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using vaes. *arXiv preprint arXiv:1807.03653*, 2018.

[139] P. Nemenyi. *Distribution-free Multiple Comparisons*. Princeton University, 1963.

[140] Hao Ni, Lukasz Szpruch, Magnus Wiese, Shujian Liao, and Baoren Xiao. Conditional sig-wasserstein gans for time series generation, 2020.

[141] A. Nica and R. Speicher. *Lectures on the Combinatorics of Free Probability*. Cambridge University Press, 2006.

[142] Ivan V Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.

[143] Pierre Pansu. Metriques de Carnot-Caratheodory et Quasiisometries des Espaces Symetriques de rang un. *Annals of Mathematics*, 129(1):1–60, 1989.

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

[145] Giovanni Peccati and Murad S Taqqu. *Wiener chaos: Moments, cumulants and diagrams: A survey with computer implementation*, volume 1. Springer Science & Business Media, 2011.

[146] G. C Pflug and A Pichler. A distance for multistage stochastic optimization models. *SIAM Journal on Optimization*, 2012.

[147] G. C Pflug and A Pichler. *Multistage stochastic optimization*. Springer, 2014.

[148] G. C Pflug and A Pichler. Dynamic generation of scenario trees. *Computational Optimization and Applications*, 2015.

[149] G. C Pflug and A Pichler. From empirical observations to tree models for stochastic optimization: convergence properties. *SIAM Journal on Optimization*, 2016.

[150] A Pichler. Evaluations of risk measures for different probability measures. *SIAM Journal on Optimization*, 2013.

[151] Novi Quadrianto, Le Song, and Alex Smola. Kernelized sorting. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1289–1296, 2009.

[152] C Reutenauer. *Free Lie Algebras*. Clarendon press – Oxford, 1993.

[153] Walter Rudin. *Principles of mathematical analysis*. McGraw-Hill Book Company, Inc., New York-Toronto-London, 1953.

[154] L Rüschendorf. The wasserstein distance and approximation theorem. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 1985.

[155] R.A Ryan. Introduction to tensor products of banach spaces. *Springer*, 2002.

[156] H. Saigo and T. Hasebe. The monotone cumulants. *Ann. Inst. Henri Poincare Probab. Stat. 47, No. 4 (2011), 1160-1170*, 2010.

[157] Saburou Saitoh and Yoshihiro Sawano. *Theory of Reproducing Kernels and Applications*. Springer Singapore, 2016.

[158] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

[159] Hiroaki Sakoe and Seibi Chiba. A dynamic programming approach to continuous speech recognition. In *Proceedings of the Seventh International Congress on Acoustics, Budapest*, volume 3, pages 65–69, Budapest, 1971. Akadémiai Kiadó.

[160] Cristopher Salvi, Maud Lemercier, Chong Liu, Blanka Hovarth, Theodoros Damoulas, and Terry Lyons. Higher order kernel mean embeddings to capture filtrations of stochastic processes. *arXiv preprint arXiv:2109.03582*, 2021.

[161] Leonard J Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

[162] Patrick Schäfer and Ulf Leser. Multivariate time series classification with weasel+muse. *ArXiv*, abs/1711.11343, 2017.

[163] Alexander Schell and Harald Oberhauser. Nonlinear independent component analysis for continuous-time signals. 2021.

[164] Bernhard Schölkopf and Alexander J. Smola. *Learning with kernels : support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. MIT Press, 2002.

[165] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[166] Dino Sejdinovic, Arthur Gretton, and Wicher Bergsma. A kernel test for three-variable interactions. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1124–1132, 2013.

[167] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, 41:2263–2291, 2013.

[168] Robert Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Statistics, 1980.

[169] C.-J. Simon-Gabriel, Alessandro Barp, and Lester Mackey. Metrizing weak convergence with maximum mean discrepancies. *arXiv: 2006.09268v1*, 2020.

[170] Alexander Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory (ALT)*, pages 13–31, 2007.

[171] Terry P. Speed. Cumulants and partition lattices. *Australian Journal of Statistics*, 25(2):378–388, 1983.

[172] Terry P. Speed. Cumulants and partition lattices II. *Australian Journal of Statistics*, 4(1):34–53, 1984.

[173] R. Speicher. Free probability theory and non-crossing partitions. *Séminaire Lotharingien de Combinatoire*, 1997.

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

[175] R. P Stanley. *Enumerative Combinatorics*. Cambridge University Press, 1986.

[176] Richard P Stanley. *Ordered structures and partitions*, volume 119. American Mathematical Soc., 1972.

[177] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.

[178] Ingo Steinwart, Chloé Pasin, R. C. Williamson, and Siyu Zhang. Elicitation and identification of properties. In *COLT*, 2014.

[179] T Sturm. Verbände von kernern isotoner abbildungen. *Czechoslovak Mathematical Journal*, 1971.

[180] T Sturm. On the lattices of kernels of isotonic mappings. II. *Czechoslovak Mathematical Journal*, 1977.

[181] Martin Sundermeyer, Tamer Alkhouli, Joern Wuebker, and Hermann Ney. Translation modeling with bidirectional recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 14–25, 2014.

[182] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014.

[183] Zoltán Szabó and Bharath K. Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18(233):1–29, 2018.

[184] Gábor Székely and Maria Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5:1249–1272, 2004.

[185] Gábor Székely and Maria Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93:58–80, 2005.

[186] Gábor J. Székely and Maria L. Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, 3:1236–1265, 2009.

[187] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35:2769–2794, 2007.

[188] Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J Maybank. General tensor discriminant analysis and gabor features for gait recognition. *IEEE transactions on pattern analysis and machine intelligence*, 29(10):1700–1715, 2007.

[189] Csaba Toth, Patric Bonnier, and Harald Oberhauser. Seq2tens: An efficient representation of sequences by low-rank tensor projections. In *International Conference on Learning Representations*, 2021.

[190] 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.

[191] Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[192] J. W Tukey. Keeping moment-like sampling computations simple. *The Annals of Mathematical Statistics*, 1956.

[193] Kerem Sinan Tuncel and Mustafa Gokce Baydogan. Autoregressive forests for multivariate time series modeling. *Pattern Recognition*, 73:202–215, 2018.

[194] Madeleine Udell and Alex Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 2019.

[195] A. V. Van der Waart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, 2000.

[196] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[197] Julio Backhoff Veraguas, Mathias Beiglböck, Manu Eder, and Alois Pichler. Fundamental properties of process distances. *Stochastic Processes and their Applications*, Apr 2020.

[198] A. M. Vershik. Decreasing sequences of measurable partitions and their applications. *Sov. Mat. Dokl.*, 1970.

[199] A. M. Vershik. Theory of decreasing sequences of measurable partitions. *Algebra i Analiz*, 1994.

[200] Z. Wang, W. Yan, and T. Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1578–1585, 2017.

[201] S. Weinberger. Speech accent archive. *George Mason University*, 2013.

[202] Stephen Willard. *General Topology*. Addison-Wesley, 1970.

[203] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[204] Tianlin Xu, Li K Wenliang, Michael Munn, and Beatrice Acciaio. Cot-gan: Generating sequential data via causal optimal transport. *arXiv preprint arXiv:2006.08571*, 2020.

[205] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

[206] Qibin Zhao, Masashi Sugiyama, Longhao Yuan, and Andrzej Cichocki. Learning efficient tensor representations with ring-structured networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8608–8612. IEEE, 2019.

[207] A. Zinger, A. Kakosyan, and Lev Klebanov. A characterization of distributions by mean values of statistics and certain probabilistic metrics. *Journal of Soviet Mathematics*, 1992.

[208] V. Zolotarev. Probability metrics. *Theory of Probability and its Applications*, 28:278–302, 1983.