The Annals of Applied Probability
2023, Vol. 33, No. 3, 2136–2175
https://doi.org/10.1214/22-AAP1862
This research was funded, in whole or in part, by [Swiss National Science Foundation, P2EZP2-188068]. A CC BY 4.0 license is applied to this article arising from this submission, in accordance with the grant's open access conditions.

ADAPTED TOPOLOGIES AND HIGHER RANK SIGNATURES

BY PATRIC BONNIER[1,a], CHONG LIU[2,c] AND HARALD OBERHAUSER[1,b]

[1]*Mathematical Institute, University of Oxford,* [a]*patric.bonnier@maths.ox.ac.uk,* [b]*harald.oberhauser@maths.ox.ac.uk*
[2]*Institute of Mathematical Sciences, ShanghaiTech University,* [c]*liuchong@shanghaitech.edu.cn*

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.

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

$$\int_{\mathbb{R}^d} f(x)\mathbb{P}_n(X_n \in dx) \to \int_{\mathbb{R}^d} f(x)\mathbb{P}(X \in dx) \quad \text{for all } f \in C_b\big(\mathbb{R}^d, \mathbb{R}\big).$$

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 1.1.

EXAMPLE 1.1 (Aldous (1981), Backhoff-Veraguas et al. (2020)).

1. The value map of an optimal stopping problem,

$$\mathbf{X} := \big(\Omega, (\mathcal{F}_t)_{t\in I}, \mathbb{P}, (X_t)_{t\in I}\big) \mapsto \inf_{\gamma} \mathbb{E}[L_\gamma], \tag{1}$$

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 (1) is restricted to the space of discrete-time Markov processes equipped with their natural filtration.

2. Figure 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 (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.

Received April 2021; revised April 2022.
MSC2020 subject classifications. Primary 60G07; secondary 60L10.
Key words and phrases. Adapted topologies, signatures, extended weak convergence.

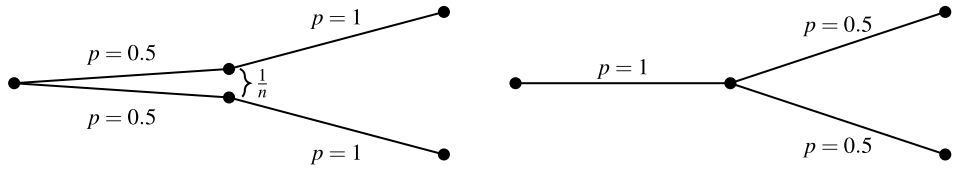


FIG. 1. *A typical example of when weak convergence is insufficient. The process on the left can be made arbitrarily close to the process on the right as* $n \to \infty$.

1.1. *Adapted topologies.* Such shortcomings of weak convergence for stochastic processes were recognized and addressed in the 1970s and 1980s. Denote by $\mathcal{S}$ the set of adapted processes that evolve in a state space that is a 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 (i.e., weak convergence in the space of measure-valued processes). Aldous studied this topology in Aldous (1981) and showed that it has several attractive properties such as making the map in Example 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 Backhoff et al. (2022), Eder (2019), Lassalle (2018), Rüschendorf (1985), Veraguas et al. (2020), Vershik (1994), Veršik (1970) including ones in other communities such as economics Hellwig (1996), operations research Pflug and Pichler (2012), Pflug and Pichler (2014), Pflug and Pichler (2015), Pflug and Pichler (2016), Pichler (2013), and numerics Bion-Nadal and Talay (2019) 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 Backhoff-Veraguas et al. (2020). 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 1.2

EXAMPLE 1.2 (Example 3.2, Hoover and Keisler (1984)). 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 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 A for details.

Seminal work of Hoover and Keisler (1984) 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 Hoover and Keisler (1984) for the precise statement. We refer to τ_r as the *adapted topology of rank* r on $\mathcal{S}$. The approach of Hoover and Keisler (1984) is different than Aldous' approach that relies on prediction processes. The starting point of Hoover and Keisler (1984) is that one may specify a

[1] Aldous refers to this topology as the *extended weak topology*.

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

$$\mathbf{X}_n \to \mathbf{X}, \quad \text{if and only if} \quad \int_{(\mathbb{R}^d)^I} f(x)\mathbb{P}_n(X_n \in dx) \to \int_{(\mathbb{R}^d)^I} f(x)\mathbb{P}(X \in dx),$$

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, for example, $f(\mathbf{X})(\omega) = \mathbb{E}[\cos(X_{t_1} X_{t_2})|\mathcal{F}_{t_3}](\omega)$ 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 built with r nested conditional expectations. Hoover–Keisler showed that by defining

$$\mathbf{X}_n \to \mathbf{X} \quad \text{if and only if} \quad \mathbb{E}\big[f(\mathbf{X}_n)\big]^2 \to \mathbb{E}\big[f(\mathbf{X})\big] \quad \text{for all } f \in \mathrm{AF}_r,$$

then the associated topologies get strictly finer as $r \to \infty$; for example, the topology τ_2 separates the two adapted processes in Example 1.2.

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

$$\mathcal{S} \times \mathcal{S} \to [0, \infty), \quad (\mathbf{X}, \mathbf{Y}) \mapsto \big\|\Phi_r(\mathbf{X}) - \Phi_r(\mathbf{Y})\big\| \tag{2}$$

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 Chevyrev and Oberhauser (2018) that the expected signature map Φ_0 can metrize weak convergence; for $r = 1$ this adds a novel entry to the

[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.

[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 $\mathbf{T}$.

list of semi-metrics that induce τ_1, see Backhoff-Veraguas et al. (2020); 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 (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 respectively the semi-metric (2) can be efficiently computed for Markov processes.

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 5 we give a simple supervised classification example that demonstrates expected signatures as they are currently used in machine learning, that is, 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 Xu et al. (2020), and the use of Φ_r in this context seems to be an interesting future research venue.

REMARK 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; for example, 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 Chevyrev and Friz (2019), Friz and Shekhar (2017) 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 4.3 for details.

REMARK 2. Our results are not restricted to stochastic processes evolving in compact subsets of finite-dimensional state spaces discussed above. By using robust signatures Chevyrev and Oberhauser (2018) adapted processes that evolve in general separable Banach space that are included in our approach. In this noncompact case, it turns out that the Hoover–Keisler approach of specifying an adapted topology via AF_r and the natural generalization of Aldous' approach given by iterating prediction process yield in general different topologies which might be of independent interest.

[4]Analogous, to how the moment sequence of a vector-valued random variable yields a metric for measures on $\mathbb{R}^d$; see Section 3.1.2. We also draw attention to Bartl, Beiglböck and Pammer (2021).

1.3. *Outline and notation.* The rest of the paper is laid out as follows:

- Section 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
$$\hat{X}^{r+1}_t := \mathbb{P}(\hat{X}^r \in \cdot | \mathcal{F}_t), \quad \hat{X}^0 := X.$$
These prediction processes evolve in state spaces that have a rich structure; for example,
$$\mathrm{Law}(\hat{\mathbf{X}}^0) \in \mathrm{Meas}(I \to V) =: \mathcal{M}_1,$$
$$\mathrm{Law}(\hat{\mathbf{X}}^1) \in \mathrm{Meas}\big(I \to \mathrm{Meas}(I \to V)\big) =: \mathcal{M}_2,$$
$$\mathrm{Law}(\hat{\mathbf{X}}^3) \in \mathrm{Meas}\big(I \to \mathrm{Meas}\big(I \to \mathrm{Meas}(I \to V)\big)\big) =: \mathcal{M}_3.$$
We refer to the spaces $\mathcal{M}_r$ as *rank r measures*. Capturing their structure is the central theme of this article.
- Section 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 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
$$\Phi_r : \mathcal{S} \to \mathbf{T}^{r+1}, \quad \Phi_r(\mathbf{X}) := \bar{\mathrm{S}}_{r+1}\big(\mathrm{Law}(\hat{X}^r)\big)$$
induces τ_r as initial topology on $\mathcal{S}$.
- Section 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 A contains details for Example 1.2, Appendix B contains some details on the construction of higher rank tensor algebras, and Appendix B.4 contains some background on the robust signature and how it can be used to overcome problems arising from non-compactness.

2. The adapted topology τ_r and the extended weak topology $\hat{\tau}_r$. In this section we recall work of Hoover–Keisler Hoover and Keisler (1984), and define adapted functionals AF_r of rank r. We then revisit Aldous' Aldous (1981) 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

[5] Available at https://github.com/PatricBonnier/Higher-rank-signature-regression

TABLE 1
List of symbols

Symbol	Meaning	Page
	Spaces	
V	a separable Banach space	2144
U	a topological space	2141
$\mathcal{S}(\mathrm{U})$	the set of adapted stochastic processes in U	2141
$\underline{\Omega}$	an adapted probability space $\underline{\Omega} = (\Omega, \mathbb{P}, (\mathcal{F}_t))$	2141
$\mathbf{X} = (\underline{\Omega}, X) \in \mathcal{S}(\mathrm{U})$	an adapted process on the stochastic base $\underline{\Omega}$	2143
$\mathrm{Meas}(\mathrm{U})$	Borel measures on U	2144
$\mathrm{Prob}(\mathrm{U})$	Borel probability measures on U	2144
I	A finite totally ordered set (time)	2136
$(I \to \mathrm{U})$	the space of sequences in U indexed by I	2146
	The Adapted Topology of Rank r	
AF	adapted functionals, $f(\mathbf{X})$ is a real-valued random variable	2138
$\mathrm{AF}_r = \{f \in \mathrm{AF} \mid \mathrm{rank}(f) \le r\}$	adapted functionals with rank less than r	2138
τ_r	the adapted topology of rank r on $\mathcal{S}(\mathrm{U})$	2
$\hat{\tau}_r$	the extended weak topology of rank r on $\mathcal{S}(\mathrm{U})$	1
	Paths and Measures of Rank r	
U_r	the space of rank r paths with state space U	2147
$\mathcal{M}_r(\mathrm{U})$	the space of rank r Borel measures on U	2149
$\mathcal{P}_r(\mathrm{U})$	the space of rank r Borel probability measures on U	2149
	(Expected) Signature of Rank r	
$\mathbf{T}^r(V)$	The rank r tensor algebra	2148
$\hat{\mathbf{X}}^r$	the rank r-prediction process $\hat{\mathbf{X}}^r_t := \mathbb{P}(\hat{\mathbf{X}}^{r-1} \in \cdot \mid \mathcal{F}_t)$ of $\mathbf{X} \in \mathcal{S}$	2143
S_r	the rank r signature map $\mathrm{S}_r : V_r \to \mathbf{T}^r(V)$	2148
$\bar{\mathrm{S}}_r$	the rank r expected signature map $\bar{\mathrm{S}}_r : \mathcal{M}_r(V) \to \mathbf{T}^r(V)$	2149
$\bar{\mathbf{X}}^r$	the rank r conditional expected signature $\bar{\mathbf{X}}^r_t := \mathbb{E}[\bar{\mathbf{X}}^{r-1} \mid \mathcal{F}_t]$	2152
$d_r(\mathbf{X}, \mathbf{Y})$	the rank r adapted signature distance between $\mathbf{X}$ and $\mathbf{Y}$	2151

functionals of rank r respectively 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 1. 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,

$$\mathcal{S}(\mathrm{U}) := \{\mathbf{X} \mid \mathbf{X} \text{ is an adapted stochastic process indexed by } I \text{ that evolves in U}\}.$$

We also set

$$\mathrm{Law}(\mathbf{X}) := \mathbb{P} \circ X^{-1} \quad \text{for } \mathbf{X} = \big(\Omega, \mathbb{P}, (\mathcal{F})_{t \in I}, X\big).$$

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.

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

$$\mathbf{X}_n \to \mathbf{X} \quad \text{if } \mathbb{E}\big[f(\mathbf{X}_n)\big] \to \mathbb{E}\big[f(\mathbf{X})\big] \quad \text{as } n \to \infty$$

for every f in this set of functionals. By choosing the set of functionals to be

$$\big\{f \mid f(\mathbf{X}) = g(X_{t_1}, \dots, X_{t_n}),\ g \in C_b\big(\mathrm{U}^n, \mathbb{R}\big),\ (t_1, \dots, t_n) \in I^n\big\}$$

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 2. We define a set of maps AF from $\mathcal{S}(\mathrm{U})$ into the set of real-valued random variables inductively:

1. if $t_1, \dots, t_n \in I$ and $f \in C_b(\mathrm{U}^n, \mathbb{R})$, then $\mathbf{X} \mapsto f(X(t_1), \dots, X(t_n)) \in \mathrm{AF}$,
2. if $f_1, \dots, f_n \in \mathrm{AF}$ and $f \in C_b(\mathbb{R}^n, \mathbb{R})$, then $\mathbf{X} \mapsto f(f_1(\mathbf{X}), \dots, f_n(\mathbf{X})) \in \mathrm{AF}$,
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 3. 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 $\mathbf{X}$.

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 1 shows two processes that can not be distinguished without at least one iteration, and in Example 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 3. Define $\mathrm{rank} : \mathrm{AF} \to \mathbb{N} \cup \{0\}$ as:

1. $\mathrm{rank}(f) = 0$ if $f(\mathbf{X}) = g(X_{t_1}, \dots, X_{t_n})$ for $g \in C_b(\mathrm{U}^n, \mathbb{R})$
2. $\mathrm{rank}(f) = \max(\mathrm{rank}(f_1), \dots, \mathrm{rank}(f_n))$ if $f(\mathbf{X}) = g(f_1(\mathbf{X}), \dots, f_n(\mathbf{X}))$, $g \in C_b(\mathbb{R}^n, \mathbb{R})$, $f_1, \dots, f_n \in \mathrm{AF}$,
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

$$\mathrm{AF}_r := \big\{f \in \mathrm{AF} \mid \mathrm{rank}(f) \le r\big\}$$

the set of adapted functionals of rank less than r.

REMARK 4. Following Definition 2, 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, $f(\mathbf{X}) = g(X_{t_1}, \dots, X_{t_n}) = \mathbb{E}[g(X_{t_1}, \dots, X_{t_n})|\mathcal{F}_T]$ 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 3 is more intuitive.

[6] In Hoover and Keisler (1984) AF are called conditional processes.

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 4. 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

$$\hat{X}^0 := X \quad \text{and} \quad \hat{X}^{r+1} := \big(\mathbb{P}(\hat{X}^r \in \cdot | \mathcal{F}_t)\big)_{t\in I}.$$

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

$$\hat{X}^r_0 = \mathrm{Law}(\hat{X}^{r-1}).$$

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 5. 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

$$\mathbb{E}\big[f(\mathbf{X})\big] = \mathbb{E}\big[f(\mathbf{Y})\big] \quad \forall f \in \mathrm{AF}_r.$$

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

$$\lim_{n\to\infty} \mathbb{E}\big[f(\hat{\mathbf{X}}^{r,n})\big] = \mathbb{E}\big[f(\hat{\mathbf{X}}^r)\big] \quad \forall f \in C_b(\mathrm{U}_r, \mathbb{R})$$

where U_r denotes the state space of process $\hat{\mathbf{X}}^r$.

2. the adapted topology of rank r if

$$\lim_{n\to\infty} \mathbb{E}\big[f(\mathbf{X}_n)\big] = \mathbb{E}\big[f(\mathbf{X})\big] \quad \forall f \in \mathrm{AF}_r.$$

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 we show that

$$\big(\mathcal{S}(\mathrm{U}), \tau_r\big) = \big(\mathcal{S}(\mathrm{U}), \hat{\tau}_r\big)$$

whenever U is compact but that for noncompact subsets U of Banach spaces, τ_r is in general coarser than $\hat{\tau}_r$; that is, $\tau_r \subsetneq \hat{\tau}_r$.

3. (Expected) signatures of rank r. In the previous Section 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 noncompact). 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, Chevyrev and

Oberhauser (2018). 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.

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.

3.1.1. *Moments and duality.* Recall that for any compact set $\mathrm{U} \subseteq V = \mathbb{R}^d$, the moment map

$$\mathrm{Meas}(\mathrm{U}) \hookrightarrow \prod_{m\geq 0} V^{\otimes m}, \quad \mu \mapsto \left(\int x^{\otimes m} \mu(dx) \right)_{m\geq 0} \tag{3}$$

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 (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 (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 noncompact spaces, see Chevyrev and Oberhauser (2018). 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 $\mathrm{Prob}(\mathrm{U}) \subset \mathrm{Meas}(\mathrm{U})$, the injection (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 needs to be taken in the noncompact case and we return to this discussion in Section 3.5.

3.1.2. *Tensors*, *moments*, *exponentials*. The tensor exponential provides a concise, coordinate-free way of expressing moment relations.

DEFINITION 6. 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 $\mathrm{U} \subset \mathbb{R}^d$, all the moments of X,

$$\mathbb{E}[\exp X] = \left(1, \mathbb{E}[X], \frac{1}{2!}\mathbb{E}[X^{\otimes 2}], \dots\right)_{m\geq 0} \in \prod_{m\geq 0} V^{\otimes m} \tag{4}$$

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 1. *Let (X_n) be a sequence of random variables that take values in a compact subset $\mathrm{U} \subset \mathbb{R}^d$. Then X_n converges weakly to a random variable X if and only if*

$$\mathbb{E}[\exp X_n] \to \mathbb{E}[\exp X] \quad \text{as } n \to \infty$$

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

$$\mathbb{E}\big[p(X)\big] = \mathbb{E}\big[p(Y)\big]$$

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

$$(\Omega, \mathcal{F}, X) \mapsto \varphi\big((\Omega, \mathcal{F}, X)\big) := \left(\mathbb{E}\left[\frac{X^{\otimes m}}{m!}\right]\right)_{m\geq 0} \in \prod_{m\geq 0} V^{\otimes m}, \tag{5}$$

respectively the weak topology is induced by the (semi-)metric

$$(\Omega_1, \mathcal{F}, X) \times (\Omega_2, \mathcal{G}, Y) \mapsto \big\|\varphi\big((\Omega_1, \mathcal{F}, X)\big) - \varphi\big((\Omega_2, \mathcal{G}, Y)\big)\big\|.$$

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 (Lang ((2012), Chapter 16)). So far we have not used this algebra structure, but we are about derive the analogous statement to (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.

3.2. *Signatures as noncommutative 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 noncommutative product.

DEFINITION 7. For $s = (s_m)_{m\geq 0}, t = (t_m)_{m\geq 0} \in \prod_{m\geq 0} V^{\otimes m}$ define

$$s \cdot t := \left(\sum_{i=0}^{m} s_i t_{m-i}\right)_{m\geq 0} \in \prod_{m\geq 0} V^{\otimes m}. \tag{6}$$

We refer to $s \cdot t$ as as the so-called tensor convolution product of s and t.

To account for 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 (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

$$\Delta_t x := \big(t+1, x(t+1)\big) - \big(t, x(t)\big) = \big(1, x(t+1) - x(t)\big) \in \mathbb{R} \oplus V$$

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 B and C for details on $\mathbf{T}^1(V)$.

Putting everything together results in the definition of the (discrete time) signature.

DEFINITION 8. Let V be a Banach space and $I = \{0, 1, \ldots, T\}$. The (rank 1) signature map is defined as

$$\mathrm{S} : (I \to V) \to \mathbf{T}^1(V), \quad x \mapsto \prod_{t\in I} \exp \Delta_t x,$$

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; respectively 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),

$$\mathrm{S}(x_1, x_0) = \exp(\Delta x) = \exp\big((x_1 - x_0, 1)\big) = 1 + \Delta x + \frac{1}{2}(\Delta x)^{\otimes 2} + \frac{1}{6}(\Delta x)^{\otimes 3} + \cdots.$$

From this point of view, the following theorem is then not surprising.

THEOREM 1. *Let V be a Banach space and $\mathrm{U} \subset V$ compact.*

1. *The map $\mathrm{S} : (I \to V) \to \mathbf{T}^1(V)$ is injective.*
2. *The family of linear signature functionals*

$$\left\{ x \mapsto \langle l, \mathrm{S}(x) \rangle : \ l \in \bigoplus_{m \geq 0} (V^{\otimes m})^\star \right\}$$

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 (Chen ((1958), Theorem 1)). Item 2 is due to Fliess (Fliess ((1976), 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 5. Everything in this section is classical: our discrete signature coincides with Chen's Chen (1954) 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 un-parametrized paths; see Chevyrev and Oberhauser (2018) for a discussion for the case of weak convergence, $r = 0$. See also Appendix C for more on signatures.

3.3. *Paths and signatures of higher rank.* Section 3.2 recalled that the [expected] signature can characterize [measures on] paths. Our goal is to characterize the predictions processes introduced in Section 2. Simply applying the expected signature to a prediction process would ignore the nested structure of the state spaces of these processes, see Definition 4, and we heavily use this structure in the proof of our main result, Section 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 9. 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,

$$\mathrm{U}_r := (I_r \to \mathrm{U}_{r-1}).$$

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

$$\mathrm{U}_r = I_r \to \mathrm{U}_{r-1} = \big(I_r \to \big(I_{r-1} \to \cdots \underbrace{\big(I_2 \to \underbrace{(I_1 \to \mathrm{U})}_{\mathrm{U}_1}\big)}_{\mathrm{U}_2} \cdots\big)\big)}_{\mathrm{U}_{r-1}}.$$

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 $t_r \in I_r$.

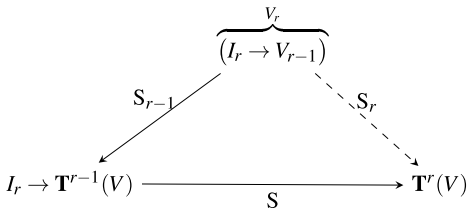


FIG. 2. *The inductive definition of* S_r. *By extending the map* S_{r-1} *to a map* $V_r \to (I_r \to \mathbf{T}^{r-1}(V))$, *the signature* S *can be applied to it to form* $\mathrm{S}_r : V_r \to \mathbf{T}^r(V)$.

Recall that the signature from Definition 8 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 B for details) we get the following.

DEFINITION 10. 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 (see Figure 2) by setting $\mathrm{S}_1 := \mathrm{S}$ and for $r \geq 2$,

$$\mathrm{S}_r(x) := \mathrm{S}\big(x^\star \mathrm{S}_{r-1}\big),$$

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.

EXAMPLE 3.1. Schematically, we can think of rank r signatures and rank r paths as

$$V_r = \big(I_r \to \big(I_{r-1} \to \cdots \big(I_2 \to \underbrace{\underbrace{\underbrace{(I_1 \to V)}_{V_1 \hookrightarrow \mathbf{T}^1(V)}}_{V_2 \hookrightarrow \mathbf{T}^2(V)}}_{V_{r-1} \hookrightarrow \mathbf{T}^{r-1}(V)}\big)\cdots\big)\big).$$

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

$$x^\star \mathrm{S}_1 : I_2 \to \mathbf{T}^1(V), \quad t_2 \mapsto \mathrm{S}_1\big(x(t_2)\big).$$

[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 $\mathbf{T}^1(V)$,

$$\mathrm{S}_2(x) = \mathrm{S}\big(x^\star \mathrm{S}_1\big),$$

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 2. *The map* $\mathrm{S}_r : V_r \to \mathbf{T}^r(V)$ *is injective.*

PROOF. Follows by iterating Theorem 1. □

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,

$$\begin{aligned}
&\mathrm{Law}\big(\hat{\mathbf{X}}^0\big) \in \mathrm{Meas}(I \to V) =: \mathcal{M}_1,\\
&\mathrm{Law}\big(\hat{\mathbf{X}}^1\big) \in \mathrm{Meas}\big(I \to \mathrm{Meas}(I \to V)\big) =: \mathcal{M}_2,\\
&\mathrm{Law}\big(\hat{\mathbf{X}}^2\big) \in \mathrm{Meas}\big(I \to \mathrm{Meas}\big(I \to \mathrm{Meas}(I \to V)\big)\big) =: \mathcal{M}_3.
\end{aligned}$$

Capturing this nested structure is essential when discussing the adapted topologies.

DEFINITION 11. Let $I_1, \dots, I_r$ be finite ordered sets and U a topological space. For $r = 1$ define $\mathcal{M}_0 := \mathrm{Prob}_0 := \mathrm{U}$ and for $r \geq 1$

$$\begin{aligned}
\mathcal{M}_r(\mathrm{U}) &:= \mathrm{Meas}\big(I_r \to \mathcal{M}_{r-1}(\mathrm{U})\big),\\
\mathcal{P}_r(\mathrm{U}) &:= \mathrm{Prob}\big(I_r \to \mathcal{P}_{r-1}(\mathrm{U})\big).
\end{aligned}$$

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

$$\begin{aligned}
\mathcal{M}_r(\mathrm{U}) &= \mathrm{Meas}\big(I_r \to \mathrm{Meas}\big(I_{r-1} \to \cdots \mathrm{Meas}\big(I_2 \to \mathrm{Meas}(I_1 \to \mathrm{U})\big)\cdots\big)\big),\\
\mathcal{P}_r(\mathrm{U}) &= \mathrm{Prob}\big(I_r \to \mathrm{Prob}\big(I_{r-1} \to \cdots \mathrm{Prob}\big(I_2 \to \mathrm{Prob}(I_1 \to V)\big)\cdots\big)\big).
\end{aligned} \tag{7}$$

As mentioned in Section 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 (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 12. 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)

$$\bar{\mathrm{S}}_r : \mathcal{M}_r(V) \to \mathbf{T}^r(V), \quad \mu \mapsto \int \mathrm{S}\big(x^\star \bar{\mathrm{S}}_{r-1}\big)\mu(\mathrm{d}x),$$

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 3 generalizes that expected signature characterizes laws of processes (Theorem 1 item 3). We postpone its proof to Theorem 4.

PROPOSITION 3. *Let V be a separable Banach space and $\mathrm{K} \subset V$ compact. Then*

$$\bar{\mathrm{S}}_r : \mathcal{P}_r(\mathrm{K}) \to \mathbf{T}^r(V)$$

is injective.

EXAMPLE 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 $((I \to \mathcal{M}_{r-1}(V)), \mu)$, 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 $\mu \in \mathcal{M}_2(V) = \mathcal{M}(I \to \mathcal{M}_1(V))$, 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

$$\bar{\mathrm{S}}_2(\mu) = \mathbb{E}\big[S\big(t \mapsto \mathbb{E}\big[S(s \mapsto Z_s)|\mathcal{F}_t\big]\big)\big].$$

We will give a complete description of $\bar{\mathrm{S}}_r$ for this special case in Section 4.

3.5. *Noncompactness 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 noncompact; in particular, Proposition 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 Chevyrev and Oberhauser (2018) yields an extension of the injectivity of the higher rank expected signature from the previous sections to paths in general (noncompact) Banach spaces. We emphasize that the results in Section 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. *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.

[8]Here the superscript "n" of S_r^n means "normalized" and does not stand for the "nth level of signature" which was often used in some literature.

PROOF. The fact that every space $\mathbf{T}^r(V)$ for $r \geq 1$ is a separable Banach space follows immediately from Definition 18. 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 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 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

$$\bar{\mathrm{S}}_r^n(\mu) = \mathbb{E}_{Z^{r-1} \sim \mu}\big[\mathrm{S}_r^n(Z^{r-1})\big]$$

with the convention $\bar{\mathrm{S}}_0^n(Z^0) = Z^0$. If $r = 1$, then this follows from Proposition 9 in Appendix 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 11. 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 13. Let V be a separable Banach space, $\mathrm{U} \subset V$, and $r \geq 0$. For the rest of this article we denote

$$\mathrm{S}_r : \mathrm{U}_r \to \mathbf{T}^r(V), \quad x \mapsto \mathrm{S}_r^n(x),$$
$$\bar{\mathrm{S}}_r : \mathcal{M}_r(\mathrm{U}) \to \mathbf{T}^r(V), \quad \mu \mapsto \bar{\mathrm{S}}_r^n(\mu).$$

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

$$\mathcal{S} \to \mathbf{T}^1(V), \quad \mathbf{X} \to \bar{\mathrm{S}}_1\big(\mathrm{Law}\big(\hat{\mathbf{X}}^0\big)\big)$$

is the topology of weak convergence $(\mathcal{S}, \tau_0)$, see Chevyrev and Oberhauser (2018). This suggests that, at least locally, the initial topology of the map

$$\mathcal{S} \to \mathbf{T}^r(V), \quad \mathbf{X} \to \bar{\mathrm{S}}_{r+1}\big(\mathrm{Law}\big(\hat{\mathbf{X}}^r\big)\big)$$

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 14. Let V be a separable Banach space and $\mathrm{U} \subset V$. For $r \geq 0$ define

$$\Phi_r : \mathcal{S}(\mathrm{U}) \to \mathbf{T}^{r+1}(V), \quad \mathbf{X} \mapsto \bar{\mathrm{S}}_{r+1}\big(\mathrm{Law}\big(\hat{\mathbf{X}}^r\big)\big)$$

and

$$d_r : \mathcal{S}(\mathrm{U}) \times \mathcal{S}(\mathrm{U}) \to [0, \infty), \quad (\mathbf{X}, \mathbf{Y}) \mapsto \big\|\Phi_r(\mathbf{X}) - \Phi_r(\mathbf{Y})\big\|_{r+1}.$$

Our main result is the following.

THEOREM 2. *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.*

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 1 (Backhoff-Veraguas et al. (2020)). *Let* U *be as in Theorem 2 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 2.

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 a useful interpretation in terms of conditional expectations; for example, for $r = 1$ and $t \in I$,

$$\bar{\mathrm{S}}_1(\hat{\mathbf{X}}^1_t) = \int \mathrm{S}(x)\mathbb{P}[X \in dx|\mathcal{F}_t] = \mathbb{E}\big[\mathrm{S}(X)|\mathcal{F}_t\big]. \tag{8}$$

This motivates the following definition.

DEFINITION 15. 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

$$\bar{X}^r_t := \mathbb{E}\big[\mathrm{S}(\bar{\mathbf{X}}^{r-1})|\mathcal{F}_t\big]$$

and $\bar{\mathbf{X}}^0_t = X_t$. We call $\bar{\mathbf{X}}^r$ the rank r conditional signature process of $\mathbf{X}$.

PROPOSITION 5. *For every* $r \geq 1$ *and* $\mathbf{X} \in \mathcal{S}(V)$ *it holds that*

$$\bar{\mathrm{S}}_r(\hat{\mathbf{X}}^r_t) = \bar{\mathbf{X}}^r_t \quad \forall t \in I. \tag{9}$$

In particular,

$$\Phi_r(\mathbf{X}) \equiv \bar{\mathrm{S}}_{r+1}\big(\mathrm{Law}(\hat{\mathbf{X}}^r)\big) = \mathbb{E}\bar{\mathbf{X}}^{r+1}_0.$$

PROOF. The second claim follows immediately from (9) since

$$\mathbb{E}\bar{\mathbf{X}}^r_t = \mathbb{E}\bar{\mathrm{S}}_r(\hat{\mathbf{X}}^r_t) = \mathbb{E}\int \bar{\mathrm{S}}_r(x)\mathbb{P}\big[\hat{\mathbf{X}}^{r-1} \in dx|\mathcal{F}_t\big] = \int \bar{\mathrm{S}}_r(x)\mathbb{P}\big[\hat{\mathbf{X}}^{r-1} \in dx\big] = \bar{\mathrm{S}}_r\big(\mathrm{Law}(\hat{\mathbf{X}}^{r-1})\big).$$

For the proof of (9) we proceed by induction over $r \geq 1$. The starting case, $r = 1$, is given in (8). For the induction step, assume that (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

$$\bar{\mathrm{S}}_{r+1}(\hat{\mathbf{X}}_t^{r+1}) = \int \mathrm{S}(x^\star \bar{\mathrm{S}}_r)\mu_r(dx) = \mathbb{E}\big[\mathrm{S}\big(s \mapsto \bar{\mathrm{S}}_r(\hat{X}_s^r)\big)|\mathcal{F}_t\big] = \mathbb{E}\big[\mathrm{S}\big(s \mapsto \bar{\mathbf{X}}_s^r\big)|\mathcal{F}_t\big],$$

where we used the induction hypothesis, $\bar{\mathrm{S}}_r(\hat{X}_s^r) = \mathbf{X}_s^r$ 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 3.

4.2. *Embedding and metrizing adapted topologies*.

THEOREM 3. *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. $\mathbb{E}[f(\mathbf{X})] = \mathbb{E}[f(\mathbf{Y})]\ \forall f \in \mathrm{AF}_r$,
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 3 with a lemma.

LEMMA 1. 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 1. 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 $B = B_0 \times \cdots \times B_T$, where each B_i is a Borel set of $(I \to \mathcal{M}_{r-1}(V))$. Hence we have

$$1_B \circ \hat{\mathbf{X}}^r = 1_{B_0} \circ \hat{\mathbf{X}}_0^r \times \cdots \times 1_{B_T} \circ \hat{\mathbf{X}}_T^r.$$

Furthermore, since $\mathrm{Meas}(I \to \mathcal{M}_{r-1}(V))$ carries the Borel σ–algebra generated by the sets of the form

$$e_U^{-1}(J) := \big\{\mu \in \mathrm{Meas}\big(I \to \mathcal{M}_{r-1}(V)\big) : e_U(\mu) := \mu(U) \in J\big\},$$
$$U \text{ Borel set in } \big(I \to \mathcal{M}_{r-1}(V)\big), \quad J \subseteq [0,1],$$

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 $\hat{\mathbf{X}}_t^r = \mathbb{P}(\hat{\mathbf{X}}^{r-1} \in \cdot | \mathcal{F}_t)$, it holds that for all t,

$$1_{B_t} \circ \hat{\mathbf{X}}_t^r = 1_{J_t} \circ \mathbb{E}\big[1_{U_n} \circ \hat{\mathbf{X}}^{r-1} | \mathcal{F}_t\big].$$

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 $1_{J_0}\times\cdots\times 1_{J_l}$ pointwise, so that $1_B\circ\hat{\mathbf{X}}^r=\lim_{j\to\infty}\varphi_j(\hat{\mathbf{X}}_0^r,\dots,\hat{\mathbf{X}}_T^r)$ a.s. (up to taking a subsequence if necessary). From the above observations we see that for each j,

$$\varphi_j(\hat{\mathbf{X}}_0^r,\dots,\hat{\mathbf{X}}_T^r)=\lim_{k\to\infty}\varphi_j\big((\mathbb{E}[f_k^n(\mathbf{X})|\mathcal{F}_t]\big)_{t\in I}),$$

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 3. $1\implies 2$. Using Lemma 1, 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

$$\mathbb{E}\big[1_B\circ\hat{\mathbf{X}}^r\big]=\mathbb{E}\big[1_B\circ\hat{\mathbf{Y}}^r\big]\quad\text{for any Borel set } B$$

that is, $\mathrm{Law}(\hat{\mathbf{X}}^r)=\mathrm{Law}(\hat{\mathbf{Y}}^r)$.

$2\implies 3$. For a Polish space $\mathcal{X}$, let $\mathrm{Meas}_a(\mathcal{X})\subset\mathrm{Meas}(\mathcal{X})$ be the set of Dirac measures on $\mathcal{X}$ $\{\delta_x : x\in\mathcal{X}\}$. Define $p:\mathrm{Meas}_a(\mathcal{X})\to\mathcal{X}$ by $p(\delta_x):=x$ and note that p is continuous with respect to the subspace topology on $\mathrm{Meas}_a(\mathcal{X})$. Define

$$\pi_T:(I\to\mathcal{X})\to\mathcal{X},\qquad \pi_T(x)=x_T\quad\text{for } x=(x_1,\dots,x_T)\in(I\to\mathcal{X}),\quad I=\{1,\dots,T\}.$$

For $g:\mathcal{X}\to\mathcal{X}$ define $\mathrm{id}_{\mathcal{X}}\oplus g:\mathcal{X}\to\mathcal{X}^2$, as $(\mathrm{id}_{\mathcal{X}}\oplus g)(x)=(x,g(x))$. 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

$$\hat{\mathbf{X}}_T^i=\delta_{\hat{\mathbf{X}}^{i-1}}\in\mathrm{Meas}_a(I\to\mathrm{Meas}_{r-1}).$$

For each r, define

$$g_r:I\to\mathrm{Meas}_r,\quad g_r=p\circ\pi_T.$$

Using that $g_r(\hat{\mathbf{X}}^r)=\hat{\mathbf{X}}^{r-1}$, it follows that for $r\geq 1$,

$$(\hat{\mathbf{X}}^r,\dots,\hat{\mathbf{X}}^{r-s})=(\mathrm{id}\oplus g_{r-s+1})\circ(\hat{\mathbf{X}}^r,\dots,\hat{\mathbf{X}}^{r-s+1}),$$

where id is applied to $\mathcal{X}=\mathrm{Meas}_r^I\times\cdots\times\mathrm{Meas}_1^I$. Since, $\mathrm{id}\oplus g_r$ is continuous we can iterate this composition to build a continuous function G, such that

$$(\hat{\mathbf{X}}^r,\dots,\hat{\mathbf{X}}^0)=G(\hat{\mathbf{X}}^r).$$

As a result, for any bounded continuous function F defined on $\mathrm{Meas}_r^I\times\cdots\times\mathrm{Meas}_0^I$, we have

$$\mathbb{E}[F(\hat{\mathbf{X}}^r,\dots,\hat{\mathbf{X}}^0)]=\mathbb{E}[F\circ G(\hat{\mathbf{X}}^r)].$$

Using 2 and denoting for brevity

$$E:=(I\to\mathrm{Meas}_r(\mathcal{X}))\times\cdots\times(I\to\mathrm{Meas}_1(\mathcal{X}))$$

we deduce that

$$\begin{aligned}\mathrm{Law}(\hat{\mathbf{X}}^r)=\mathrm{Law}(\hat{\mathbf{Y}}^r)\quad&\Rightarrow\quad\forall F\in C_b(E),\quad\mathbb{E}[F\circ G(\hat{\mathbf{X}}^r)]=\mathbb{E}[F\circ G(\hat{\mathbf{Y}}^r)]\\&\Leftrightarrow\quad\forall F\in C_b(E),\quad\mathbb{E}[F(\hat{\mathbf{X}}^r,\dots,\hat{\mathbf{X}}^0)]=\mathbb{E}[F(\hat{\mathbf{Y}}^r,\dots,\hat{\mathbf{Y}}^0)]\\&\Leftrightarrow\quad\mathrm{Law}(\hat{\mathbf{X}}^0,\dots,\hat{\mathbf{X}}^r)=\mathrm{Law}(\hat{\mathbf{Y}}^0,\dots\hat{\mathbf{Y}}^r).\end{aligned}$$

$3 \implies 1$. We prove by induction that for any $r \geq 0$, and $f \in \mathrm{AF}_r$, there exists some bounded and Borel measurable $\tilde{f} : (I \to \mathrm{Meas}_r(V)) \to \mathbb{R}$

$$f(\mathbf{X}) = \tilde{f}(\hat{\mathbf{X}}^r).$$

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 $f \in \mathrm{AF}_{r-1}$, there exists a bounded Borel measurable function $\tilde{f}$ defined on $(I \to \mathrm{Meas}_{r-1}(V))$ 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 $\hat{\mathbf{X}}^r_t = \mathbb{P}(\hat{\mathbf{X}}^{r-1} \in \cdot|\mathcal{F}_t)$ is the regular conditional distribution of $\hat{\mathbf{X}}^{r-1}$ given $\mathcal{F}_t$, we also obtain that

$$\mathbb{E}\big[\tilde{f}(\hat{\mathbf{X}}^{r-1})|\mathcal{F}_t\big] = e_{\tilde{f}}(\pi_t \circ \hat{\mathbf{X}}^r),$$

where π_t is the tth coordinate mapping such that $\pi_t \circ \hat{\mathbf{X}}^r = \hat{\mathbf{X}}^r_t$, and $e_{\tilde{f}}$ is the evaluation map defined on $\mathrm{Meas}_r(V) = \mathrm{Meas}(I \to \mathrm{Meas}_{r-1}(V))$ such that $e_{\tilde{f}}(\mu) := \int \tilde{f} d\mu$. Since $\tilde{f}$ is bounded and measurable by (Bertsekas and Shreve ((1978), Corollary 7.29.1)), $e_{\tilde{f}}$ is a bounded measurable function on $\mathrm{Meas}_r(V)$.

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, $\bar{\mathrm{S}}_r$ is injective on $\mathrm{Prob}_r(V)$ hence the equivalence follows immediately from Proposition 5 and the fact that $\mathrm{Law}(\hat{\mathbf{X}}^r), \mathrm{Law}(\hat{\mathbf{Y}}^r) \in \mathrm{Prob}_{r+1}(V)$. □

The metrics d_r (cf. Definition 14) locally characterize the rank r extended weak topology $\hat{\tau}_r$.

PROPOSITION 6. *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 5, $\bar{\mathbb{E}}[\mathbf{X}_0^{r+1}] = \mathbb{E}[\mathrm{S} \circ \bar{\mathrm{S}}_r(\hat{\mathbf{X}}^r)]$, and by Proposition 4, $\mathrm{S} \circ \bar{\mathrm{S}}_r$ 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, $\|\mathbb{E}\bar{\mathbf{X}}_0^{r+1} - \mathbb{E}\bar{\mathbf{Y}}_0^{r+1}\|_{r+1} = 0$. Now using Theorem 3 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 5).

PROPOSITION 7. *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 $\mathrm{K} \subset V$, the converse holds, that is,*

$$(\mathcal{S}(\mathrm{K}), \tau_r) = (\mathcal{S}(\mathrm{K}), \hat{\tau}_r).$$

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 $\lim_{k\to\infty} \mathbb{E}[f(\hat{\mathbf{X}}^{k,r})] = \mathbb{E}[f(\hat{\mathbf{X}}^r)]$; that is, $\mathbf{X}^k$ 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 2) one can easily verify that the class $\mathcal{A} := \{\rho_f : f \in \mathrm{AF}_r\}$ is a subalgebra in $C_b(I \to \mathcal{P}_r(\mathrm{K}); \mathbb{R})$. Moreover, using the proof of $1 \implies 2$ in Theorem 3 we can also prove that $\mathcal{A}$ separate points on $(I \to \mathcal{P}_r(\mathrm{K}))$. Therefore, since the space $(I \to \mathcal{P}_r(\mathrm{K}))$ is obviously compact (recall that $\mathrm{Prob}(\mathrm{K})$ is compact in the weak topology, and then by induction one can prove the compactness for all $\mathcal{P}_r(\mathrm{K}))$, $\mathcal{A}$ is dense in $C_b(K; \mathbb{R})$ under the uniform topology by the Stone–Weierstrass theorem. Hence, for any given $\rho \in C_b(I \to \mathcal{P}_r(\mathrm{K}); \mathbb{R})$ and any $\varepsilon > 0$, we can pick a $\rho_f \in \mathcal{A}$ such that $\sup_{x\in(I\to\mathcal{P}_r(\mathrm{K}))} |\rho(x) - \rho_f(x)| \leq \varepsilon$, 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 which in turn implies Theorem 2.

THEOREM 4. *Let V be a separable Banach space and $r \geq 0$. Then for $\mathbf{X}, \mathbf{Y} \in \mathcal{S}(V)$*

$$\big(\mathbb{E}[f(\mathbf{X})] = \mathbb{E}[f(\mathbf{Y})] \quad \forall f \in \mathrm{AF}_r\big) \text{ if and only if } \Phi_r(\mathbf{X}) = \Phi_r(\mathbf{Y}).$$

Moreover:

1. *the map Φ_r locally induces the topology τ_r,*
2. *the semimetric d_r locally metrizes the topology τ_r.*

If $\mathrm{K} \subset V$ is compact, then the above statements apply without localization, as stated in Theorem 2. 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 Chevyrev and Oberhauser (2018); restricted to $r = 1$ this (locally) induces Aldous extended weak topology Aldous (1981). With $r = 1$ and $(\mathcal{F}_t)_{0\leq t\leq T}$ the natural filtration it adds another entry to the list in Backhoff-Veraguas et al. (2020) of distances that induce the same topology, and therefore we obtain Corollary 1.

REMARK 6. A natural question is that of quantitative guarantees and put simply, we do not have any nontrivial 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 Backhoff-Veraguas et al. (2020), Bartl, Beiglböck and Pammer (2021) shows that the value function

$$v : \mathbf{X} = \big(\Omega, (\mathcal{F}_t)_{t\in I}, \mathbb{P}, (X_t)_{t\in I}\big) \mapsto \inf_{\gamma} \mathbb{E}[L_\tau]$$

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 Backhoff et al. (2022) for the causal Wasserstein distance, seems yet another promising research direction.

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 8. *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 C, see Proposition 12 and Corollary 2.

Now we consider $r = 1$; that is, 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 2. A set $\mathrm{U} \subset \mathcal{P}_2(\mathbb{R}^d)$ is tight if and only if the set

$$\{F_\sharp \mu : \mu \in \mathrm{U}\} \subset \mathcal{P}_1\big(\mathbf{T}^1\big(\mathbb{R}^d\big)\big)$$

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 8 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 2 we obtain the following characterization of tightness set for the extended weak topology:

THEOREM 5. *Let* $\mathrm{U} \subset \mathcal{S}(\mathbb{R}^d)$.

1. U *is tight in Aldous' extended weak topology if and only if the collection of martingale laws*

$$\big\{\mathrm{Law}\big(\big(\mathbb{E}\big[\mathrm{S}(X)|\mathcal{F}_t\big]\big)_{t\in I}\big) : \mathbf{X} \in \mathrm{U}\big\}$$

is tight in $\mathrm{Prob}(I \to \mathbf{T}^1(\mathbb{R}^d))$.

2. *If in addition all filtrations are natural and* U *is closed in Aldous' extended weak topology, then the following are equivalent*:

- U *is compact in Aldous' extended weak topology, Hellwig's information topology, and all other topologies mentioned in Corollary* 1.
- U *satisfies Eder's conditions* (*Eder* ((2019), *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 7.

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 8. This observation also prevents 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 (Bartl, Beiglböck and Pammer ((2021), 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 8, 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 5 complements Eder's tightness theorem (Eder ((2019), 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, for example, 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 $\mathbf{T}^1(\mathbb{R}^d)$;

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 3. 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

$$u_t(a) = \mathbb{E}[Z_{t+1} \cdots Z_T | Z_t = a] = \mathbb{E}\big[Z_{t+1} u_{t+1}(Z_{t+1}) | Z_t = a\big]. \qquad \square$$

[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.

Algorithm 1 Pseudo-code for $\Phi_0(\mathbf{X})$

```
Input: A Markov chain X represented as a rooted tree with root r
procedure ExpSig_0(a)
    if a.children is empty then
        return 1
    sum ← 0
    for b in a.children do
        sum ← sum + p(a, b) · exp{b − a} · ExpSig_0(b)
    return sum
Output: ExpSig_0(r) = Φ_0(X)
```

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

$$Z_t := \exp(\Delta_t X) \quad \text{with } \Delta_t X := (X_{t+1} - X_t, 1)$$

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 3 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

$$u_t(x) = \sum_{y \in V} \exp(y - x) \mathbb{P}(X_{t+1} = y | X_t = x) u_{t+1}(y). \tag{10}$$

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 $p(a, b)$.

LEMMA 4. 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 (i.e., $\mathbf{X}$ terminates at time d) then Algorithm 1 computes $\Phi_1(\mathbf{X})$ with complexity

$$O(tN) \quad \text{in time} \quad \text{and} \quad O(ds) \quad \text{in space},$$

where t, s are the time and space costs of computing and storing one call of $\exp(\cdot)$ to the desired accuracy.

PROOF. Note that the bounds are clearly true if the tree has 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 $i = 1, \ldots, n$. 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 ith child are $O(tn_i)$ and $O(sd_i)$ respectively, hence the total time complexity is

$$O\left(\sum_{i=1}^{n} tn_i\right) = O(tN).$$

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

$$O\Big(1+s+\max_{1\le i\le n}(sd_i)\Big)=O\big(1+s(d+1)\big)=O(ds),$$

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 3 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

$$v_t^2(x)=\sum_{y\in V}\mathbb{P}(X_{t+1}=y|X_t=x)\exp\big\{\mathbb{E}\big[\mathrm{S}(X)|X_{t+1}=y\big]-\mathbb{E}\big[\mathrm{S}(X)|X_t=x\big]\big\}v_{t+1}^2(y).$$

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:

$$\begin{aligned}v_t^2(x)&=\sum_{y\in V}\mathbb{P}(X_{t+1}=y|X_t=x)\\&\qquad\times\exp\big\{\mathrm{S}(X|X_t=y)v_{t+1}^1(y)-\mathrm{S}(X|X_{t-1}=x)v_t^1(x)\big\}v_{t+1}^2(y),\\v_t^1(x)&=\sum_{y\in V}\mathbb{P}(X_{t+1}=y|X_t=x)\exp\{y-x\}v_{t+1}^1(y),\end{aligned}$$

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 (10). 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 noncommutative 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

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: $\mathrm{ExpSig}_1(r)=\Phi_1(\mathbf{X})$

one call of $\exp_1()$ to the desired accuracy, and d is the depth of x. This can be remedied by memoising the values of $S(x)$ once computed, which brings the time complexity down to $O(t)$ but takes up more space.

LEMMA 5. 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

$$O((t+T)N) \quad \text{in time} \quad \text{and} \quad O(d(MS+s)) \quad \text{in space.}$$

PROOF. The same arguments made in the proof of Lemma 4 apply here too. If one caches $S(X|X_t = a)$ at every node one needs to store at most d values of $S(\cdot)$ and the cost of computing $\exp\{S(X|X_t = b)v^1_{t+1}(b) - S(X|X_{t-1} = a)v^1_t(a)\}$ 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 8 (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 B for a more thorough discussion of how to write down the gradings, and the dimensions of the spaces involved, but note to always write down (formal) gradings $\mathbf{T}^r(V) = \prod_{k\geq 0} \mathbf{T}^r(V)_k$, where if V has dimension d, then

$$\begin{aligned}
&\dim.\mathbf{T}^1(V)_k = (d+1)^k,\\
&\dim.\mathbf{T}^2(V)_0 = 1, \quad \dim.\mathbf{T}^2(V)_1 = d+1,\\
&\dim.\mathbf{T}^2(V)_k = (2d+3)\dim.\mathbf{T}^2(V)_{k-1} - (d+1)\dim.\mathbf{T}^2(V)_{k-2}.
\end{aligned}$$

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 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 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

$$\begin{aligned}
\mathbf{X}^c&: \quad X_0 = 0, \quad X_1 = N_1, && X_2 = c + N_2,\\
\mathbf{Y}^c&: \quad Y_0 = 0, \quad Y_1 = \sqrt{1-\varepsilon^2}M_1 + \varepsilon c, && Y_2 = c + M_2,
\end{aligned}$$

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, $\mathrm{Law}(\mathbf{X}^c) \approx \mathrm{Law}(\mathbf{Y}^c)$, analogously to Example 1.1 respectively Figure 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 3 shows for each sample from μ (an adapted process) one sample trajectory from this process.

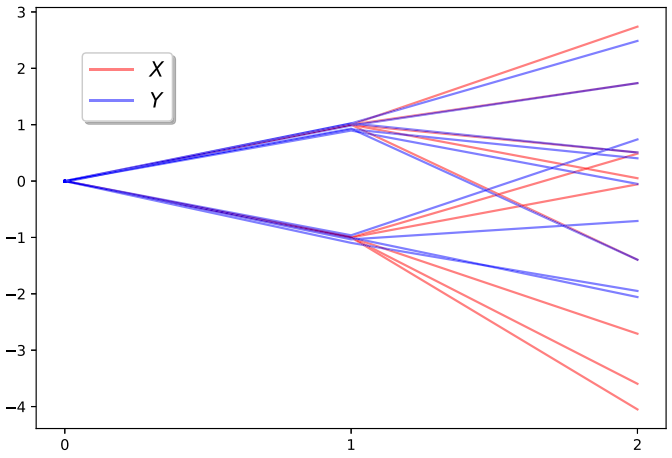


FIG. 3. *The sample paths of μ are plotted above for $\varepsilon = 0.05$. They are marked as either $\boldsymbol{X}$ in red, or $\boldsymbol{Y}$ in blue depending on if the sample comes from $\mathbf{X}^c$ or $\mathbf{Y}^c$.*

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 Hearst (1998) was trained on m data points from the training set with Φ_0 respectively Φ_1 as feature map.

Figure 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.

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

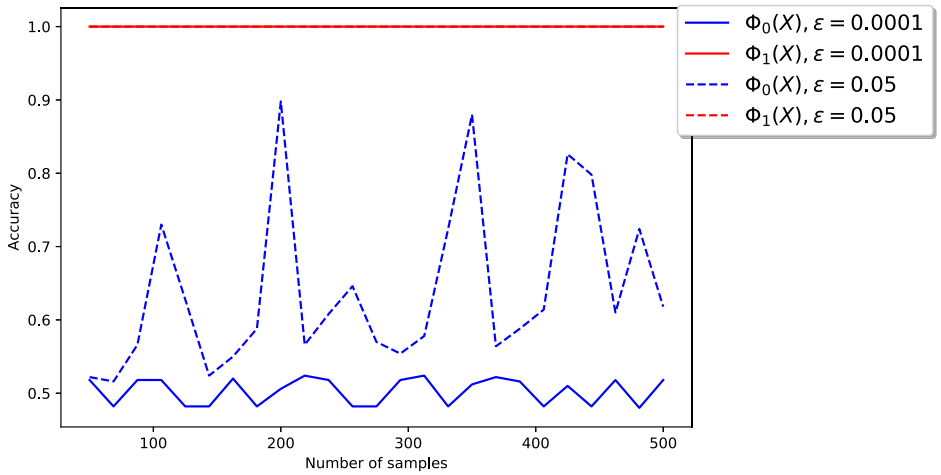


FIG. 4. *The accuracies of the linear classifier trained on $\Phi_0(\mathbf{X})$ and $\Phi_1(\mathbf{X})$ is plotted against the number of samples used in blue and red respectively. Solid lines are used for $\varepsilon = 10^{-4}$ and dashed lines for $\varepsilon = 5 \times 10^{-2}$.*

[10]This corresponds to 127 coordinates for Φ_0 resp. 76 coordinates for Φ_1, hence is in favour of Φ_0.

signatures provide additional features that linearise complex dependencies between law and filtration by representing them on multi-graded spaces of tensors.

APPENDIX A: DETAILS FOR EXAMPLE 1.2

Consider the probability space $\Omega = \{1, \dots, 16\}$ equipped with the counting measure and the filtration

$$\begin{aligned}
\mathcal{F}_0 &= \{\Omega, \varnothing\},\\
\mathcal{F}_1 &= \sigma\big\langle \{1, \dots, 8\}, \{9, \dots, 16\}\big\rangle,\\
\mathcal{F}_2 &= \sigma\big\langle \{1,2,3,4\}, \{5,6,7,8\}, \{9,10,11,12\}, \{13,14,15,16\}\big\rangle,\\
\mathcal{F}_3 &= \sigma\big\langle \{1,2\}, \{3,4\}, \{5,6\}, \{7,8\}, \{9,10\}, \{11,12\}\{13,14\}, \{15,16\}\big\rangle,\\
\mathcal{F}_4 &= 2^{\Omega}.
\end{aligned}$$

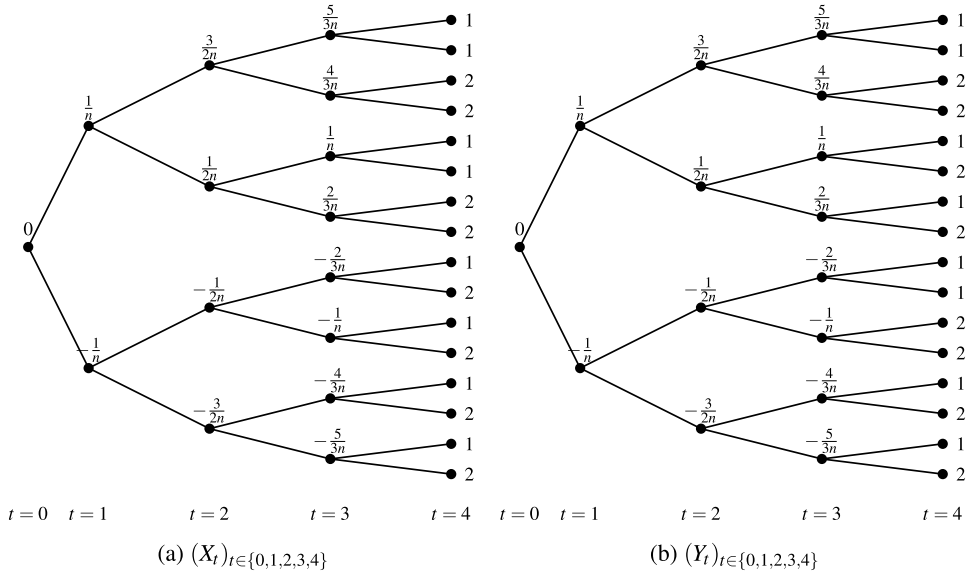
Define the two processes

$$X_0 = X_1 = X_2 = X_3 = 0, \qquad X_4 : \begin{cases} 1,2,5,6,9,11,13,15 \mapsto 1,\\ 3,4,7,8,10,12,14,16 \mapsto 2, \end{cases}$$

$$Y_0 = Y_1 = Y_2 = Y_3 = 0, \qquad Y_4 : \begin{cases} 1,2,5,7,9,10,13,15 \mapsto 1,\\ 3,4,6,8,11,12,14,16 \mapsto 2, \end{cases} .$$

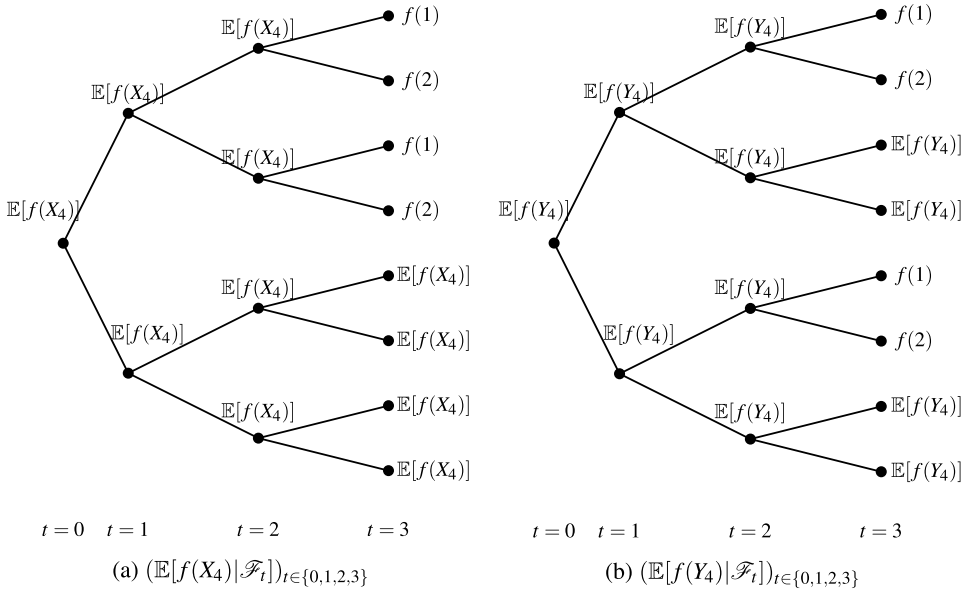
If the above construction looks unnatural the reader is also invited to think of the filtration as being the natural filtration associated to the processes and that instead of staying at 0 until time 4, they move with step size of order $1/n$ in such a way to generate $\mathcal{F}$, as in Figure 5 and 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:

$$\mathbb{E}\big[f(X_4)|\mathcal{F}_0\big] = \mathbb{E}\big[f(X_4)|\mathcal{F}_1\big] = \mathbb{E}\big[f(X_4)|\mathcal{F}_2\big] = \frac{1}{2}\big(f(1) + f(2)\big),$$



(a) $(X_t)_{t\in\{0,1,2,3,4\}}$ (b) $(Y_t)_{t\in\{0,1,2,3,4\}}$

FIG. 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 stays at 0 until $t = 4$ when it jumps to either 1 or 2.*



(a) $(\mathbb{E}[f(X_4)|\mathcal{F}_t])_{t\in\{0,1,2,3\}}$ (b) $(\mathbb{E}[f(Y_4)|\mathcal{F}_t])_{t\in\{0,1,2,3\}}$

FIG. 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.*

$$\mathbb{E}\big[f(X_4)|\mathcal{F}_3\big]: \begin{cases} \{1,2\},\{5,6\} \mapsto f(1), \\ \{3,4\},\{7,8\} \mapsto f(2), \\ \{9,10\},\{11,12\},\{13,14\},\{15,16\} \mapsto \dfrac{1}{2}\big(f(1)+f(2)\big), \end{cases}$$

$$\mathbb{E}\big[f(Y_4)|\mathcal{F}_0\big] = \mathbb{E}\big[f(Y_4)|\mathcal{F}_1\big] = \mathbb{E}\big[f(Y_4)|\mathcal{F}_2\big] = \frac{1}{2}\big(f(1)+f(2)\big),$$

$$\mathbb{E}\big[f(Y_4)|\mathcal{F}_3\big]: \begin{cases} \{1,2\},\{9,10\} \mapsto f(1), \\ \{3,4\},\{11,12\} \mapsto f(2), \\ \{5,6\},\{7,8\},\{13,14\},\{15,16\} \mapsto \dfrac{1}{2}\big(f(1)+f(2)\big), \end{cases}$$

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

$$\mathbb{E}\big[\mathbb{E}[X_4|\mathcal{F}_3]^2|\mathcal{F}_1\big]: \begin{cases} \{1,2,3,4,5,6,7,8\} \mapsto \dfrac{5}{2}, \\ \{9,10,11,12,13,14,15,16\} \mapsto \dfrac{9}{4}, \end{cases}$$

$$\mathbb{E}\big[\mathbb{E}[Y_4|\mathcal{F}_3]^2|\mathcal{F}_1\big] = \frac{19}{8}.$$

Hence the information structure in these processes are different, but this can't be seen by their prediction processes alone.

APPENDIX 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 16. 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 Ryan (2002) 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 nonunital tensor algebra over V. The higher order tensor algebras are defined inductively as follows:

DEFINITION 17. Let V be a normed space. Define the spaces
$$\mathbf{ut}^0(V) = V, \qquad \mathbf{ut}^r(V) = \bigoplus_{m\geq 0} \left(\mathbf{ut}^{r-1}(V) - 1\right)^{\otimes m}, \quad r \geq 1.$$
$$\mathbf{t}^0(V) = V, \qquad \mathbf{t}^r(V) = \bigoplus_{m\geq 0} \left(\mathbb{R} \oplus \mathbf{t}^{r-1}(V) - 1\right)^{\otimes m}, \quad r \geq 1.$$

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 $\mathcal{F}[\cdot]$ for the free algebra generated by $\cdot$ and $\mathcal{M}[\cdot]$ 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 6. Let $A_1,\dots,A_n$ be multi-graded modules with respective multi-gradings $M_1,\dots,M_n$. Then $\mathcal{F}[A_1,\dots,A_n]$ is multi-graded by $\mathcal{M}[M_1,\dots,M_n]$.

Using the notation $\otimes_{(r)}$ for the tensor product on $\mathbf{ut}^{r-1}(V)$, we note that by the above lemma $(\mathbf{ut}^r(V), +, \otimes_{(r)})$ 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
$$\mathbf{ut}^1(V)_{\mathbf{k}} := V^{\otimes_{(1)}\mathbf{k}}, \qquad \mathbf{ut}^r(V) = \bigoplus_{\mathbf{k}\in \mathrm{Seq}^r} \mathbf{ut}^r(V)_{\mathbf{k}},$$
$$\text{where } \mathbf{ut}^r(V)_{\mathbf{k}} := \mathbf{ut}^{r-1}(V)_{\mathbf{k}_1} \otimes_{(r)} \cdots \otimes_{(r)} \mathbf{ut}^{r-1}(V)_{\mathbf{k}_l} \quad \text{for } \mathbf{k} = \mathbf{k}_1 \cdots \mathbf{k}_l \in \mathrm{Seq}^r.$$

We also use the following recursive definition of the degree for a multi-index:
$$\deg.\mathbf{k} = \deg.\mathbf{k}_1 + \cdots + \deg.\mathbf{k}_l, \quad \text{for } \mathbf{k} = \mathbf{k}_1 \cdots \mathbf{k}_l \in \mathrm{Seq}^r, \quad \deg.\varnothing = 1.$$

Which allows us to write down a grading for $\mathbf{ut}^r(V)$ as
$$\mathbf{ut}^r(V) = \bigoplus_{k\geq 0} \Big(\bigoplus_{\substack{\mathbf{k}\in \mathrm{Seq}^r, \\ \deg.\mathbf{k}=k}} \mathbf{ut}^r(V)_{\mathbf{k}} \Big). \tag{11}$$

See (Ebrahimi-Fard and Patras ((2015), Section 3)) for more on $\mathbf{ut}^1(V)$ and $\mathbf{ut}^2(V)$.

EXAMPLE 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
$$\mathbf{ut}^2(V) = 1 + \bigoplus_{n_1,\dots,n_k \geq 1} V^{\otimes_{(1)} n_1} \otimes_{(2)} \cdots \otimes_{(2)} V^{\otimes_{(1)} n_k}.$$
- $\mathbf{ut}^3(V)$ is graded over matrices in $\mathbb{N}$ by
$$\mathbf{ut}^3(V) = 1 + \bigoplus_{\substack{n_1^1,\dots,n_{k_1}^1 \geq 1 \\ \vdots \\ n_1^{k_2},\dots,n_{k_1}^{k_2} \geq 1}} \left(V^{\otimes_{(1)} n_1^1} \otimes_{(2)} \cdots \otimes_{(2)} V^{\otimes_{(1)} n_{k_1}^1}\right) \otimes_{(3)} \cdots \otimes_{(3)} \left(V^{\otimes_{(1)} n_1^{k_2}} \otimes_{(2)} \cdots \otimes_{(2)} V^{\otimes_{(1)} n_{k_1}^{k_2}}\right).$$

REMARK 9. 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 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,
$$\begin{aligned}\kappa(X_1, X_2, X_3) &= \mathbb{E}(X_1X_2X_3) - \mathbb{E}(X_1)\mathbb{E}(X_2X_3) - \mathbb{E}(X_2)\mathbb{E}(X_1X_3) \\ &\quad - \mathbb{E}(X_3)\mathbb{E}(X_1X_2) + 2\mathbb{E}(X_1)\mathbb{E}(X_2)\mathbb{E}(X_3) \\ &= \mu_X^\star(e_1e_2e_3) - \mu_X^\star(e_1 \otimes_{(2)} e_2e_3) - \mu_X^\star(e_2 \otimes_{(2)} e_1e_3) \\ &\quad - \mu_X^\star(e_3 \otimes_{(2)} e_1e_2) + 2\mu_X^\star(e_1 \otimes_{(2)} e_2 \otimes_{(2)} e_3).\end{aligned}$$

B.1. The multi-grading of $\mathbf{t}^r(V)$. Recall that the signature map, Definition 8, 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 6 it is multi-graded by $\mathcal{M}(\mathcal{M}^1, \mathcal{M}^0)$. We may write:

$$\mathbf{t}^2(V) = \bigoplus_{m_1,\dots,m_n \in \mathcal{M}^1, \mathcal{M}^0} \mathbf{t}^2(V)_{m_1} \otimes_{(2)} \cdots \otimes_{(2)} \mathbf{t}^2(V)_{m_n},$$

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 (11) 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.

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 (11)):

$$\mathbf{ut}^1(V) = \bigoplus_{k\geq 0} \mathbf{ut}^1(V)_k, \qquad \dim . \mathbf{ut}^1(V)_k = d^k.$$

In the case of $\mathbf{ut}^2(V)$ we may define $\mathbf{ut}^2(V)_k := \bigoplus_{\mathbf{k}\in \mathrm{Seq}^2, \deg.\mathbf{k}=k} \mathbf{ut}^2(V)_{\mathbf{k}}$ and write

$$\mathbf{ut}^2(V) = \bigoplus_{k\geq 0} \mathbf{ut}^2(V)_k, \qquad \dim . \mathbf{ut}^2(V)_k = \frac{1}{2}(2d)^k.$$

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

$$\mathbf{t}^1(V) = \bigoplus_{k\geq 0} \mathbf{t}^1(V)_k, \qquad \dim . \mathbf{t}^1(V)_k = (d+1)^k.$$

The case $\mathbf{t}^2(V)$ is slightly more complicated, but can be characterised by a simple linear recursion.

PROPOSITION 6. *Let V be a d-dimensional vector space, then*

$$\mathbf{t}^2(V) = \bigoplus_{k\geq 0} \mathbf{t}^2(V)_k, \qquad \dim . \mathbf{t}^2(V)_k := A_{d+1}(k),$$

where A_d satisfies the recursion

$$A_d(k) = (2d+1)A_d(k-1) - d\,A_d(k-2), \qquad A_d(0) = 1, \qquad A_d(1) = d+1.$$

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 $k_{i_1}, \dots, k_{i_l} \in \mathcal{M}^1$, then since $\deg k_i = 1$ for any $i \neq i_1, \dots, i_l$ it must be true that $\deg . \mathbf{k} = \sum_{k_i \in \mathcal{M}^1} k_i + n - l$. By setting $\mathbf{t}^2(V)_k := \bigoplus_{\mathbf{k}\in\mathcal{M}(\mathcal{M}^1,\mathcal{M}^0), \deg.\mathbf{k}=k} \mathbf{t}^2(V)_{\mathbf{k}}$ we see that

$$\dim \mathbf{t}^2(V)_k = \sum_{n=0}^{k} \sum_{l=0}^{n} \#\big\{\mathbf{k} = k_1 \cdots k_n \in \mathcal{M}\big(\mathcal{M}^1, \mathcal{M}^0\big) : \deg . \mathbf{k} = k,$$
$$k_{i_1}, \dots, k_{i_l} \in \mathcal{M}^1\big\}(d+1)^{l+k-n}.$$

We note that for n, l fixed we have

$$\begin{aligned}&\#\{\mathbf{k} = k_1 \cdots k_n \in \mathcal{M}(\mathcal{M}^1, \mathcal{M}^0) : \deg_{\cdot}(\mathbf{k}) = k, k_{i_1}, \ldots, k_{i_l} \in \mathcal{M}^1\}\\ &= \binom{n}{l} \#\{\mathbf{k} = k_1 \cdots k_l \in \mathcal{M}^2 : \deg_{\cdot}(\mathbf{k}) = k + l - n\}\\ &= \binom{n}{l}\binom{k+l-n-1}{k-n}.\end{aligned}$$

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 $k \geq 2$

$$\begin{aligned}&\dim \mathbf{t}^2(V)_k - (2d+3)\dim \mathbf{t}^2(V)_{k-1} + (d+1)\dim \mathbf{t}^2(V)_{k-2}\\ &= \sum_{n=2}^{k-1}(d+1)^n\big[(k+1-n)\,{}_2F_1(1-n, k+2-n, 2, -1)\\ &\quad - 2(k+1-n)\,{}_2F_1(2-n, k+2-n, 2, -1)\\ &\quad + (k-n)\,{}_2F_1(2-n, k+1-n, 2, -1) - (k-n)\,{}_2F_1(1-n, k+1-n, 2, -1)\big]\\ &\quad + (d+1)^k\big[{}_2F_1(1-k, 2, 2, -1) - 2\,{}_2F_1(2-k, 2, 2, -1)\big]\\ &\quad + (d+1)\big[k\,{}_2F_1(0, k+1, 2, -1) - (k-1)\,{}_2F_1(0, k, 2, -1) - 1\big].\end{aligned}$$

Because of the two facts

$${}_2F_1(-k, 2, 2, -1) = 2^k, \qquad {}_2F_1(0, k, 2, -1) = 1,$$

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

$$\begin{aligned}f(a+1, b, k) &= \frac{a+1}{a+1-k}f(a, b, k), \qquad f(a, b+1, k) = \frac{b+k}{b}f(a, b, k),\\ f(a, b, k+1) &= \frac{(a-k)(b+k)}{(k+1)(k+2)}f(a, b, k).\end{aligned}$$

By iterating these three relations one can show that

$$(b+1)f(a,b+2,k)-2(b+1)f(a-1,b+2,k)+bf(a-1,b+1,k)-bf(a,b+1,k)$$
$$=\frac{k(k+1)}{a}f(a,b+1,k)-\frac{(k+1)(k+2)}{a}f(a,b+1,k+1),$$

and the claimed recursion follows by summing over k since

$$\begin{aligned}&(b+1)F(-a,b+2)-2(b+1)F(1-a,b+2)+bF(1-a,b+1)-bF(-a,b+1)\\&=\sum_{k=-\infty}^{\infty}(b+1)f(a,b+2,k)-2(b+1)f(a-1,b+2,k)\\&\quad+bf(a-1,b+1,k)-bf(a,b+1,k)\\&=\sum_{k=-\infty}^{\infty}\frac{k(k+1)}{a}f(a,b+1,k)-\sum_{k=-\infty}^{\infty}\frac{(k+1)(k+2)}{a}f(a,b+1,k+1)=0.\end{aligned}$$

□

REMARK 10. The sequences $A_0(k)$, $A_1(k)$ are listed as $A001519$ and $A052984$ respectively on OEIS.

B.3. Higher rank tensor algebras on Banach spaces.

DEFINITION 18. 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 11. 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 12. Note that S_r indeed takes values in $\mathbf{T}^r(E)$, since by the above Remark 11 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 18 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 19. 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.

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 20. 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 (Chevyrev and Oberhauser ((2018), Proposition A.2 and Corollary A.3)).

THEOREM 7. *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 (Lyons and Qian ((2002), 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 21. We call $\mathrm{S}_r^n := \Lambda\,\mathrm{S}^r$ the robust (or, normalized) signature map of rank r and $\bar{\mathrm{S}}_r^n := \mathbb{E}\,\mathrm{S}_r^n$ is called the robust expected signature map of rank r.

The proof of the next proposition can be found in (Chevyrev and Oberhauser ((2018), Corollary 5.7)).

PROPOSITION 9. *Let V be a separable Banach space. Then $\bar{\mathrm{S}}_1^n : \mathrm{Meas}(I \to V) \to \mathbf{T}^1(V)$ is injective.*

For our concrete purpose in Section 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 10. *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. Lyons and Qian ((2002), 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 (Friz and Victoir ((2010), 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 $cT = T$, that is, $c = 1$. Consequently it follows that $\mathrm{S}(x) = \mathrm{S}(y)$ and also $x = y$ by the injecvitity of S (see Theorem 1). □

REMARK 13. Note that the injectivity of Φ depends crucially on the fact that we include time component into the definition of the 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 Chevyrev and Oberhauser (2018). Also note that we include $\|x\|_\infty$ into the dilation for a special technical reason, see discussion in the next section.

APPENDIX 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 noncompactness, see Section 3.5. Central to our argument is to exploit a duality between functions and measures via a "universal feature map". In the noncompact case this duality can be subtle to handle, see Simon-Gabriel, Barp and Mackey (2020) for an overview.

C.1. Universality and characteristicness.

DEFINITION 22. 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, for example, (Chevyrev and Oberhauser ((2018), Theorem 2.3)).

THEOREM 8. *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 $\mathcal{F}'$.*

C.2. Robust signature features and their topology. Put in our context, the feature map is the (robust) signature map S_r^n.

THEOREM 9. *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 (Chevyrev and Oberhauser ((2018), Section 5.1)). □

The following proposition is crucial for this present paper.

PROPOSITION 11. *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 9 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 $V = \mathbb{R}^d$, i.e., equivalent to the shuffle product equation in (Lyons, Caruana and Lévy ((2007), (2.6)))). This implies that $\tilde{\mathcal{F}}$ satisfies all conditions in (Chevyrev and Oberhauser ((2018), Theorem 2.6, (2))) and is therefore dense in $C_b(\mathcal{X}^I, \mathbb{R})$ with respect to the strict topology by (Chevyrev and Oberhauser ((2018), 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 (Chevyrev and Oberhauser ((2018), Theorem 2.3)). The last assertion then follows immediately. □

C.3. Kernelized maximum mean discrepancies. Following Chevyrev and Oberhauser (2018) 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 12. *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 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 10 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 $\mathbb{1} = (1, 0, \ldots) \in \mathcal{H}^1(V)$ the unit element. By (Chevyrev and Oberhauser ((2018), 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 11. 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 11, see also (Chevyrev and Oberhauser ((2018), 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

$$\begin{aligned} |\mathbf{k}(x, y)| &= |\langle \tilde{\mathrm{S}}_1^n(x), \tilde{\mathrm{S}}_1^n(y) \rangle| \leq \|\tilde{\mathrm{S}}_1^n(x)\| \|\tilde{\mathrm{S}}_1^n(y)\| \\ &= \|\tilde{\mathrm{S}}_1^n(x)\| \|\delta_{\exp(-\|\mathrm{S}(y)\| - \|y\|_\infty)} \mathrm{S}(y) - \mathbb{1}\| \\ &\leq \|\tilde{\mathrm{S}}_1^n(x)\| \exp(-\|\mathrm{S}(y)\| - \|y\|_\infty) \|\mathrm{S}(y)\| \\ &\to 0 \end{aligned}$$

as $\|y\|_\infty \to \infty$. Hence, in view of (Simon-Gabriel, Barp and Mackey ((2020), Lemma 4.1)) one can conclude that $\mathcal{H}_{\mathbf{k}} \subset C_0(I \to V)$. □

We now conclude by (Simon-Gabriel, Barp and Mackey ((2020), Lemma 2.1)).

COROLLARY 2. *Let* $V = \mathbb{R}^d$. *Then*

$$d_{\mathbf{k}}(\mu, \nu) = \left\| \int \mathrm{S}_r^n(x) \mu(dx) - \int \mathrm{S}_r^n(x) \nu(dx) \right\| = \|\bar{\mathrm{S}}_1^n(\mu) - \bar{\mathrm{S}}_1^n(\nu)\|$$

characterizes weak convergence.

Acknowledgments. HO would like to thank Manu Eder for helpful discussions. CL would like to thank Gudmund Pammer for pointing out the fact that the metric d_r characterizes the convergence in $\hat{\tau}_r$ when V is a locally compact space.

Funding. PB is supported by the Engineering and Physical Sciences Research Council [EP/R513295/1].

CL is supported by the SNSF Grant [P2EZP2_188068].

HO is supported by the EPSRC grant "Datasig" [EP/S026347/1], the Alan Turing Institute, the Oxford-Man Institute, and the Centre for Intelligent Multidimensional Data Analysis (CIMDA).

REFERENCES

ALDOUS, D. J. (1981). Weak convergence and general theory of processes. Unpublished Draft of Monograph.

BACKHOFF, J., BARTL, D., BEIGLBÖCK, M. and WIESEL, J. (2022). Estimating processes in adapted Wasserstein distance. *Ann. Appl. Probab.* **32** 529–550. MR4386535 https://doi.org/10.1214/21-aap1687

BACKHOFF-VERAGUAS, J., BARTL, D., BEIGLBÖCK, M. and EDER, M. (2020). All adapted topologies are equal. *Probab. Theory Related Fields* **178** 1125–1172. MR4168395 https://doi.org/10.1007/s00440-020-00993-8

BARTL, D., BEIGLBÖCK, M. and PAMMER, G. (2021). The Wasserstein space of filtered processes. Available at arXiv:2104.14245.

BERTSEKAS, D. P. and SHREVE, S. E. (1978). *Stochastic Optimal Control*: *The Discrete Time Case*. *Mathematics in Science and Engineering* **139**. Academic Press [Harcourt Brace Jovanovich, Publishers], New York–London. MR0511544

BION-NADAL, J. and TALAY, D. (2019). On a Wasserstein-type distance between solutions to stochastic differential equations. *Ann. Appl. Probab.* **29** 1609–1639. MR3914552 https://doi.org/10.1214/18-AAP1423

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

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

CHEVYREV, I. and FRIZ, P. K. (2019). Canonical RDEs and general semimartingales as rough paths. *Ann. Probab.* **47** 420–463. MR3909973 https://doi.org/10.1214/18-AOP1264

CHEVYREV, I. and OBERHAUSER, H. (2018). Signature moments to characterize laws of stochastic processes. Preprint. Available at arXiv:1810.10971.

EBRAHIMI-FARD, K. and PATRAS, F. (2015). Cumulants, free cumulants and half-shuffles. *Proc. A* **471** 20140843, 18. MR3325207 https://doi.org/10.1098/rspa.2014.0843

EDER, M. (2019). Compactness in adapted weak topologies.

FLIESS, M. (1976). Un Outil Algebrique: Les Series Formelles Non Commutatives. In *Mathematical Systems Theory* (G. Marchesini and S. K. Mitter, eds.) 122–148. Springer Berlin Heidelberg, Berlin, Heidelberg.

FRIZ, P. K. and SHEKHAR, A. (2017). General rough integration, Lévy rough paths and a Lévy–Kintchine-type formula. *Ann. Probab.* **45** 2707–2765. MR3693973 https://doi.org/10.1214/16-AOP1123

FRIZ, P. K. and VICTOIR, N. B. (2010). *Multidimensional Stochastic Processes as Rough Paths*: *Theory and Applications*. *Cambridge Studies in Advanced Mathematics* **120**. Cambridge Univ. Press, Cambridge. MR2604669 https://doi.org/10.1017/CBO9780511845079

HEARST, M. A. (1998). Support vector machines. *IEEE Intell. Syst.* **13** 18–28. https://doi.org/10.1109/5254.708428

HELLWIG, M. F. (1996). Sequential decisions under uncertainty and the maximum theorem. *J. Math. Econom.* **25** 443–464. MR1394333 https://doi.org/10.1016/0304-4068(95)00739-3

HOOVER, D. N. and KEISLER, H. J. (1984). Adapted probability distributions. *Trans. Amer. Math. Soc.* **286** 159–201. MR0756035 https://doi.org/10.2307/1999401

LANG, S. (2012). *Algebra* **211**. Springer, Berlin.

LASSALLE, R. (2018). Causal transport plans and their Monge–Kantorovich problems. *Stoch. Anal. Appl.* **36** 452–484. MR3784142 https://doi.org/10.1080/07362994.2017.1422747

LYONS, T. J., CARUANA, M. and LÉVY, T. (2007). *Differential Equations Driven by Rough Paths*. *Lecture Notes in Math.* **1908**. Springer, Berlin. MR2314753

LYONS, T. and QIAN, Z. (2002). *System Control and Rough Paths*. *Oxford Mathematical Monographs*. Oxford Univ. Press, Oxford. MR2036784 https://doi.org/10.1093/acprof:oso/9780198506485.001.0001

PFLUG, G. CH. and PICHLER, A. (2012). A distance for multistage stochastic optimization models. *SIAM J. Optim.* **22** 1–23. MR2902682 https://doi.org/10.1137/110825054

PFLUG, G. CH. and PICHLER, A. (2014). *Multistage Stochastic Optimization*. *Springer Series in Operations Research and Financial Engineering*. Springer, Cham. MR3288310 https://doi.org/10.1007/978-3-319-08843-3

PFLUG, G. CH. and PICHLER, A. (2015). Dynamic generation of scenario trees. *Comput. Optim. Appl.* **62** 641–668. MR3426120 https://doi.org/10.1007/s10589-015-9758-0

PFLUG, G. CH. and PICHLER, A. (2016). From empirical observations to tree models for stochastic optimization: Convergence properties. *SIAM J. Optim.* **26** 1715–1740. MR3543169 https://doi.org/10.1137/15M1043376

PICHLER, A. (2013). Evaluations of risk measures for different probability measures. *SIAM J. Optim.* **23** 530–551. MR3033118 https://doi.org/10.1137/110857088

RÜSCHENDORF, L. (1985). The Wasserstein distance and approximation theorems. *Z. Wahrsch. Verw. Gebiete* **70** 117–129. MR0795791 https://doi.org/10.1007/BF00532240

RYAN, R. A. (2002). *Introduction to Tensor Products of Banach Spaces*. *Springer Monographs in Mathematics*. Springer London, Ltd., London. MR1888309 https://doi.org/10.1007/978-1-4471-3903-4

SIMON-GABRIEL, C. J., BARP, A. and MACKEY, L. (2020). Metrizing weak convergence with maximum mean discrepancies. Available at arXiv:2006.09268v1.

VERAGUAS, J. B., BEIGLBÖCK, M., EDER, M. and PICHLER, A. (2020). Fundamental properties of process distances. *Stochastic Process. Appl.* **130** 5575–5591. MR4127339 https://doi.org/10.1016/j.spa.2020.03.017

VERSHIK, A. M. (1994). Theory of decreasing sequences of measurable partitions. *Algebra i Analiz* **6** 1–68. MR1304093

VERŠIK, A. M. (1970). Descending sequences of measurable decompositions, and their applications. *Dokl. Akad. Nauk SSSR* **193** 748–751. MR0268360

XU, T., WENLIANG, L. K., MUNN, M. and ACCIAIO, B. (2020). Cot-gan: Generating sequential data via causal optimal transport. Preprint. Available at arXiv:2006.08571.