Conditioning on Stochastic Processes via Signatures

R. Alaifari and A. Schell

Research Report No. 2023-45
December 2023

Seminar für Angewandte Mathematik
Eidgenössische Technische Hochschule
CH-8092 Zürich
Switzerland

Conditioning on Stochastic Processes via Signatures

Rima Alaifari[*]   Alexander Schell[†]

Department of Mathematics, ETH Zürich

14 December 2023

Abstract

In this working paper, the conditional expectation of a random vector given a stochastic process is characterised as the solution of some convex (semi-)infinite linear least squares problem. This result is based on a functional monotone class argument involving the robust signature of the conditioning process, and it enables the nonparametric and practically feasible computation of conditional distributions for very general classes of jointly distributed stochastic processes.

Keywords: conditional expectation, conditional distribution, conditional probability, supervised learning, nonparametric regression, functional regression, function approximation.

1 Introduction

Knowing the conditional expectation of a random variable given another random variable is fundamental to a structured understanding of the statistical dependencies between these variables. In formalising the concept of 'best-approximation' of one variable based on the information given by another, the general significance of conditional expectations extends far beyond theoretical nuances, with their practical utility permeating areas as diverse as option pricing [39] or risk assessments in financial markets and portfolios [2], stochastic filtering and the optimisation of control systems in engineering [3, 31], computer vision [1] and molecular dynamics [24], survival analysis [22] and causal models [33], time series analysis and forecasting [32, 34, 40] and Bayesian inversion and inference [30, 37] broadly, to name just a very few classical examples. Recently, one of the most spectacular domains of application for conditional expectation, in its role as the fundamental target concept behind statistical regression, is the area of statistical machine learning [17] at large and the subfield of natural language processing, most notably large language models [46], in particular, the latter seen essentially as statistical models to approximate conditional distributions[1] [38, 43].

A particular challenge in computing conditional expectations and their derived statistics, also present in most of the above-cited examples and especially in sequential or language-based machine learning, is to efficiently account for potential time-dependencies in the conditioning variable, that is: to condition on stochastic processes. This problem was first systematically considered in classical probability, where the traditional model classes of martingales and Markov processes were conceived to elegantly circumvent subtler issues of time-dependent conditioning. Yet these classical 'evading' assumptions have their clear limitations, e.g. [6, 26], and so it is worthwhile to revisit the general problem of time-dependent conditioning with modern tools from stochastic analysis.

[*] rima.alaifari@math.ethz.ch   [†] alexander.schell@math.ethz.ch

[1] A conditional distribution, of course, amounts to a family of conditional expectations over indicator functions.

This working paper[2] addresses the following related questions, which the literature suggests to be of significant relevance in practice: Given a random vector Z in $\mathbb{R}^m$ and two multiviariate stochastic processes X and Y in (discrete- or) continuous time and not necessarily of the same dimension,

$$\textit{How can the quantities} \quad \mathbb{E}[Z\,|\,X] \quad \textit{and} \quad \mathbb{P}(Y \in \cdot\,|\,X) \quad \textit{be efficiently computed?} \tag{1}$$

The approximation of conditional expectations has been well-addressed for time-independent X or if the temporality of $X = (X_t)$ conforms to certain parametric assumptions (e.g. [8, 5, 45] and the references therein), but to the best of our knowledge there are currently no rigorous nonparametric answers to (1) for general (jointly distributed) stochastic processes $X = (X_t)$ and $Y = (Y_t)$.

The present work attempts to close this gap by using tools from rough path theory, and in particular the concept of "robust signatures" recently introduced in [10], to structure the conditional σ-algebra generated by the process X: As described in Section 3, the robust signature, ϕ, is an algebraically structured and Hilbert-valued global injection over the space of sufficiently continuous paths that are the realisations of the process X. Since this injection is also bounded, a functional monotone class argument shows that the span of (already some linear subset $\mathfrak{L}$ of) all bounded linear functionals composed with $\phi(X)$ is an L^2-dense subset of all square-integrable X-measurable functions, see the proof of Proposition 4.3. This then implies as our first answer to (1) the variational characterisation:

$$\begin{gathered}\mathbb{E}[Z\,|\,X] \;=\; \lim_{k\to\infty} \sum_{i=1}^{m} \big\langle \alpha_{ik}, \phi(X) \big\rangle \cdot e_i, \quad \text{for } (\alpha_{ik})_k \text{ any minimizing sequence of} \\ \inf_{(\alpha_1,\dots,\alpha_m)\in\mathfrak{L}^m} \sum_{i=1}^{m} \mathbb{E}\big| Z^i - \langle \alpha_i, \phi(X)\rangle \big|^2,\end{gathered} \tag{2}$$

where the above convergence holds at least in L^2 and also almost surely if (α_{ik}) is fast enough, see Corollary 4.6. Noting that the optimization in (2) is convex, a similar 'convex characterisation' can be found if instead of Z given X we are interested in the conditional expectation given X of the robust signature of Y itself, see Theorem 4.5. An according characterisation of the latter then provides a variational identity of the conditional distribution of Y given X, namely, for any A Borel,

$$\mathbb{P}(Y \in A\,|\,X) \;=\; \lim_{l\to\infty} \lim_{k\to\infty} \big\langle \psi_{\alpha_k}(X), \ell_{A,l} \big\rangle, \tag{3}$$

which again holds in L^2 and also almost surely under further conditions, and where both $(\psi_{\alpha_k}(X))_k$ and $(\ell_{A,l})_l \subset \mathfrak{L}$ are explicitly computable by solving well-structured convex optimisation problems involving (observations of) (X, Y), see Section 5.1. Both (2) and (3) are practical answers to (1).

As highlighted at the beginning of this introduction, both of the above representations (2) and (3) are new algorithmic solutions to important statistical approximation problems. An identity of type (3) for sequential conditioning, in particular, is the central target property of most current machine learning architectures on sequential data, yet here achieved rigorously and on a much simpler algorithmic premise thanks to the 'linearising' quality of the signature (Propositions 4.3 and 5.1). Note further that the validity of equations (2) and (3) does not depend on assuming any prior relations between the marginals of (Z, X) or (Y, X), such as continuity or specific statistical dependence structures.

[2] While this document is technically complete and accurate to the best of our knowledge, we would like to emphasize that it is currently only a *working paper* in the sense that the presented theory is currently being further developed with a view towards adding, among others, complementary assertions on convergence rates, additional examples and illustrating applications, see also Section 6.

Closing with a note on existing literature, we remark that prior approximations in a manner similar to (2) have been first explored in [25, 27], though the respective aspects of these works are mostly empirical and based on rather strict assumptions on (Z, X). In a spirit related to ours but with no mathematical connection, [11] study nowcasting using linear regression on signatures. Finally, we note that the observation that the robust signature algebra is dense in L^p, which is crucial for (2), was also made independently of us (and by other mathematical means) in the most recent preprint [4], where this idea is taken to different consequences in the realm of optimal stopping.

This working paper is organised as follows. Section 2 presents the case for conditional distributions and their data-driven approximation (3) as a statistical foundation for supervised machine learning, to ensure that readers with no intrinsic probabilistic background or interest can appreciate the paper's main question (1) and the results we obtained. It is essentially an amateur's perspective on the role of conditioning in supervised learning and can be skipped by anyone interested only in the paper's statistical main contents (1) and (2) and (3). The latter contents are developed from Section 3 onward, where the signature transform is introduced as essentially a global coordinate map on spaces of sufficiently continuous paths (Sections 3.1 and 3.2) and then examined for its well-known but central 'linearisation' property, that even when the signature is 'squished' into a ball by tensor normalisation, continuous functions on paths still asymptotically factor through it linearly (Section 3.3). The main contribution of this working paper is then made in Section 4, where, based on a new result on the L^2-density of normalised signature algebras (Proposition 4.3), the conditional expectation of the normalised signature of a process Y given another process X is characterised as the solution of a conveniently approximable, convex semi-infinite linear least squares problem (Theorem 4.5). In addition to providing a convenient and computationally expedient variational characterisation of conditional expectations of the form (2) as a corollary to its proof, Theorem 4.5 is also seen to imply linearised asymptotic representations of the form (3) for the conditional distribution of a process Y given X with essentially no prior assumption about their joint distribution (Section 5.1). Following a brief remark on how the results of this paper are applicable to common prediction tasks (Section 5.2), the working paper concludes with a brief outlook on some of the ongoing extensions that are currently being added to it (Section 6).

2 Motivation: Conditioning to Learn Relations from Data

This section explains how the seemingly technical question (1) about conditioning and conditional distributions is of direct interest to machine learning practitioners. Note for this that many of today's challenges in machine learning involve the flexible and 'data-driven', i.e. statistical[3], modelling of input-response relationships over data from high or even infinite dimensional spaces. In other words, one cares about the 'appropriately estimated' description of (sets of) ordered pairs

$$\mathfrak{R} \ \subset \ \mathcal{X} \times \mathcal{Y} \equiv \big\{(x, y) \,\big|\, x \in \mathcal{X},\ y \in \mathcal{Y}\big\} \tag{4}$$

for certain sets $\mathcal{X}$ and $\mathcal{Y}$, which are usually vector spaces of up to infinite dimension. Traditionally, the focus has been on the data-based approximation of functional relations ('regression analysis'), i.e. on approximating from the data relations of the form

$$\mathfrak{R}_f \ = \ \big\{(x, f(x)) \,\big|\, x \in \mathcal{X}\big\} \ \cong \ f \qquad \text{for} \quad f : \mathcal{X} \to \mathcal{Y} \quad \text{some function.} \tag{5}$$

Often, these classical regression tasks rely on the a priori assumption that the 'true' relation f is in fact contained within a preparametrized family (f_θ) of well-understood proxies, from which f must then be (approximately) identified, e.g. estimated variationally as the argmin of some appropriate objective function.

[3] That is, finding approximations of the true relations that are are computable functions of the data ('estimators').

Remark 2.1. In formal terms, $\mathfrak{R}_f \in \bigcup_{\theta\in\Theta} \mathfrak{R}_f^\theta \subseteq 2^{\mathcal{X}\times\mathcal{Y}}$ for some parametrised relational base $\big(\mathfrak{R}_f^\theta := \{(x, f_\theta(x)) \mid x \in \mathcal{X}\} \mid \theta \in \Theta\big)$, where ideally the parametrisation $\theta \mapsto \mathfrak{R}_f^\theta$ is injective, so that $\mathfrak{R}_f \in \arg\min_{\theta\in\Theta} \Phi(\mathfrak{R}_f^\theta)$ for an objective $\Phi : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ such as, e.g., the square-loss $\Phi(\mathfrak{R}_f^\theta) := \int_{\mathcal{X}} \|f(x) - f_\theta(x)\|^2 \, \hat{\mu}(\mathrm{d}x)$ where $\hat{\mu} := |\mathfrak{X}|^{-1} \sum_{x\in\mathfrak{X}} \delta_x$ is the data-dependent counting measure over the data $\mathfrak{X} \subseteq \mathcal{X}$ with cardinality $|\mathfrak{X}| < \infty$. ◆

Now, for many if not most real-world situations and related machine learning tasks, the desired relations (4) do not admit a definite functional relation of the form (5). This is because, in many such cases, a given input x does not necessarily lead to a single (deterministic) output $y = f(x)$ only, but can instead be associated with a whole (i.e., generally non-singleton) set

$$\mathfrak{R}_x := \{y \in \mathcal{Y} \mid (x, y) \in \mathfrak{R}\} \tag{6}$$

of possible outcomes. (Of course, this can also be seen as a (set-valued) functional relation $\mathcal{X} \to 2^{\mathcal{Y}}$ via $x \mapsto \mathfrak{R}_x$, though this abstraction will not provide much gain for us here.) Often, however, for such multivalued associations, not all of the elementary associations y in (6) are equally 'relevant' given an input x, but some of the elements of $\mathfrak{R}_x$ *are more plausible* in relation to x *than others.*

Mathematically, one way to formalise such a 'relevance-based gradient', or 'evaluation', of an x-section $\mathfrak{R}_x$ is by means of a probability measure

$$\mu_x \,:\, \mathcal{B}(\mathcal{Y}) \to [0,1] \qquad \text{with} \qquad \mathrm{supp}(\mu_x) \subseteq \overline{\mathfrak{R}_x}\,, \tag{7}$$

where $\mathcal{B}(\mathcal{Y})$ is the Borel-σ-algebra on $\mathcal{Y}$ (or any other suitable σ-algebra to carry a measure on $\mathfrak{R}_x$). Given an input datum x, it is now a measure μ_x, instead of a single value $y = f(x)$ as in (5), that is to be learned from the data. We call $\big(\mathfrak{R}, (\mu_x \mid x \in \mathcal{X})\big)$ a *relevance-graded association (RGA).*

Example 2.2 (Machine Translation)**.** Consider the task of machine translation, where x represents a sentence in an input language $\mathcal{X}$ to be translated into a corresponding sentence y in an output language $\mathcal{Y}$: Since there is usually more than one plausible translation y for a given x, one is interested in outputs of the form (6) for the relation $(x, y) \in \mathfrak{R} \;:\Leftrightarrow\;$ "y is a valid translation of x". A plausible valuation (7) of responses (6) to x could then be a numerical validity ranking of different admissible translations of x, as compiled by polling an expert panel of professional translators. ◆

Recognising that we usually don't know exactly how to explain the data, or how it was generated, and also to account for its manifold versatility, we take a basic statistical perspective and model a data point $x \in \mathcal{X}$ as being sampled from an underlying *random variable* X, i.e. we assume that

$$x \;=\; X(\omega) \quad \text{for some} \quad \omega \in \Omega, \qquad \text{where} \qquad X \,:\, \Omega \to \mathcal{X} \quad \text{is measurable}^{4} \tag{8}$$

and defined on some fixed probability space $(\Omega, \mathscr{F}, \mathbb{P})$ (i.e., a measure space with $\mathbb{P}(\Omega) = 1$).

To connect assumption (8) with the task of learning measure-valued functions (7) of the data, we require that 'μ_x varies measurably in x', which translates to the technical assumption that the map

$$\begin{gathered} \kappa \,:\, \mathcal{X}\times\mathcal{B}(\mathcal{Y}) \to [0,1], \quad (x, A) \mapsto \mu_x(A), \\ \text{is such that:} \quad x \mapsto \kappa(x, B) \equiv \mu_x(B) \quad \text{is} \quad \big(\mathcal{B}(\mathcal{X}), \mathcal{B}([0,1])\big)\text{-measurable}, \quad \forall\, B \in \mathcal{B}(\mathcal{Y}). \end{gathered} \tag{9}$$

(Note that (9) is equivalent to stating that: the measure-valued map $\mathcal{X} \ni x \mapsto \kappa(x, \cdot) = \mu_x \in \mathcal{M}_1$ is $(\mathcal{B}(\mathcal{X}), \Sigma)$-measurable, where $\Sigma := \sigma(\mathrm{ev}_B \mid B \in \mathcal{B}(Y))$ is the σ-algebra on $\mathcal{M}_1 \equiv \mathcal{M}_1(\mathcal{B}(\mathcal{Y}))$ (the set of all Borel probability measures on $\mathcal{Y}$) that is generated by the functions $\mathrm{ev}_B : \mathcal{M}_1 \ni \mu \mapsto \mu(B)$.)

A map of the form (9) is called a *probability kernel*, and we will use this concept for an analytically clean description for the *statistical learning* of relevance-graded associative relations (4).

[4] Naturally, as part of this perspective we assume that $\mathcal{X}$ can be endowed with an associated measurable structure.

Definition 2.3. Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish. We call *regular RGA* a pair $(\mathfrak{R}, \kappa)$ for which $\mathfrak{R} \in \mathcal{B}(\mathcal{X} \times \mathcal{Y})$ is a Borel-measurable relation and $\kappa \equiv (\kappa(x, \cdot))$ is a prob. kernel with $\kappa(x, \mathfrak{R}_x) = 1$ for each $x \in \mathcal{X}$.

The statistical base model (8) extends to a natural statistical framework for learning an RGA $\big(\mathfrak{R}, (\mu_x \mid x \in \mathcal{X})\big)$ [see (4) & (7)] via the assumption that related pairs $(x, y) \in \mathfrak{R}$ can be lifted to realisations of a random pair (X, Y), with μ_x given by the conditional law of Y given $\{X = x\}$:

$$\begin{gathered}\text{there is} \quad \text{a random pair} \quad (X, Y) : \Omega \to \mathcal{X} \times \mathcal{Y} \quad \text{such that} \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y} : \\ (x, y) \in \mathfrak{R} \quad :\Leftrightarrow \quad \Big[(x, y) = (X(\omega), Y(\omega)) \ \text{ for an } \omega \in \Omega, \quad \text{and} \quad \mu_x(\cdot) = \mathbb{P}(Y \in (\cdot) \mid X = x)\Big]; \end{gathered} \tag{10}$$

here as before, $(\Omega, \mathscr{F}, \mathbb{P})$ is some probab. space carrying the marginals $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$.

For regular RGAs, the lift (10) of a relevance-graded association to a pair of $(\mathcal{X} \times \mathcal{Y})$-valued pair of random variables always exists and is essentially unique. Moreover, the converse holds as well, establishing a one-to-one correspondence between pairs of random variables and regular RGAs.

Proposition 2.4. *For each regular RGA $(\mathfrak{R}, \kappa)$ and Borel prob. measure ξ on $\mathcal{X}$ with $\xi(\overline{\pi_{\mathcal{X}}(\mathfrak{R})}) = 1$, there is a canonical pair (X, Y) of jointly distributed rvs. $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ such that*

$$\kappa(x, \cdot\,) \ = \ \mathbb{P}(Y \in (\cdot) \mid X = x) \quad \textit{for} \ \ \xi\textit{-almost every } x \in \mathcal{X}. \tag{11}$$

Conversely, each pair $(\tilde{X}, \tilde{Y}) : \Omega \to \mathcal{X} \times \mathcal{Y}$ of (jointly distributed) random variables is the lift of a regular RGA $(\tilde{\mathfrak{R}}, \tilde{\kappa})$, i.e. there is a relation $\tilde{\mathfrak{R}} \in \mathcal{B}(\mathcal{X} \times \mathcal{Y})$ and a probability kernel $\tilde{\kappa}$ such that (11) holds for $(\mathfrak{R}, \kappa, X, Y, \xi) \coloneqq (\tilde{\mathfrak{R}}, \tilde{\kappa}, \tilde{X}, \tilde{Y}, \mathbb{P}_{\tilde{X}})$, and the associated random measure $\tilde{\kappa} : \mathcal{X} \to \mathcal{M}_1(\mathcal{B}(\mathcal{Y}))$ is unique almost everywhere wrt $\mathbb{P}_{\tilde{X}}$.

Proof. This is essentially just a quote of standard results in basic probability. Indeed:

Define the measurable space $(\Omega, \mathscr{F}) \coloneqq (\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X} \times \mathcal{Y}))$ and endow it with the composition

$$\mathbb{P} \coloneqq \xi \otimes \kappa \ : \ A \ \mapsto \ \int_{\mathcal{X}} \int_{\mathcal{Y}} \mathbb{1}_A(x, y)\, \kappa(x, \mathrm{d}y)\, \xi(\mathrm{d}x) \quad \big(A \in \mathcal{B}(\mathcal{X} \times \mathcal{Y})\big).$$

Then $\mathbb{P}$ is a Borel probability measure on $\mathcal{X} \times \mathcal{Y}$, cf. [19, Lemma 3.3], and the maps

$$X \coloneqq \pi_{\mathcal{X}} \quad \text{and} \quad Y \coloneqq \pi_{\mathcal{Y}} \qquad \big(\text{hence } Z \coloneqq (X, Y) \ = \ \mathrm{id}_{\mathcal{X}} \times \mathrm{id}_{\mathcal{Y}}\big)$$

are Borel-measurable random variables on $(\Omega, \mathscr{F}, \mathbb{P})$, with $\mathbb{P}_X = \xi$ and joint law $\mathbb{P}_Z = \mathbb{P}$. The identity (11) then holds by the disintegration theorem, see for instance [19, Theorem 8.5].

For the converse statement, let $\tilde{Z} \coloneqq (\tilde{X}, \tilde{Y})$ and set $\tilde{\mathfrak{R}} \coloneqq \operatorname{supp}(\mathbb{P}_{\tilde{Z}})$ and define $\tilde{\kappa}$ as the disintegration of $\mathbb{P}_{\tilde{Z}}$ wrt $\mathbb{P}_{\tilde{X}}$, so that $\tilde{\kappa}$ is $\mathbb{P}_{\tilde{X}}$-a.e. unique with the property $\mathbb{P}_{\tilde{Z}} = \mathbb{P}_{\tilde{X}} \otimes \tilde{\kappa}$; see e.g. [19, Theorem 3.4 (i) & (ii)]. Then $\tilde{\mathfrak{R}}$ is [closed and hence] Borel, and $[\mathbb{1}_{\tilde{\mathfrak{R}}_{\tilde{X}}}(\tilde{Y}) = \mathbb{1}_{\tilde{\mathfrak{R}}}(\tilde{Z})$ a.s. and hence] $\tilde{\kappa}(\tilde{X}, \tilde{\mathfrak{R}}_{\tilde{X}}) = \mathbb{P}(\tilde{Y} \in \tilde{\mathfrak{R}}_{\tilde{X}} \mid \tilde{X}) = \mathbb{E}[\mathbb{1}_{\tilde{\mathfrak{R}}}(\tilde{Z}) \mid \tilde{X}] = 1$ a.s. [by [19, Theorem 8.5 (i)]] which implies that $\tilde{\kappa}(x, \tilde{\mathfrak{R}}_x) = 1$ for each $x \in \mathcal{X} \setminus \mathcal{N}$, for some $\mathcal{N} \subset \mathcal{X}$ with $\mathbb{P}_{\tilde{X}}(\mathcal{N}) = 0$; restrict to $\tilde{\mathfrak{R}}_x$ and normalise to $\tilde{\kappa}(x, \tilde{\mathfrak{R}}_x) = 1$ for $x \in \mathcal{N}$. The corresponding identity (11) holds by [19, Theorem 8.5 (i)]. □

In summary, the goal of supervised learning is to statistically recover binary relations (4) from data, and these relations are typically sufficiently complex to pertain (to not simply the graph of some deterministic function (5) of the input variable, as for classical curve fitting tasks, but more generally) to a random input-ouput pair (X, Y) such that the given response (6) to an input x is captured by the conditional distribution of Y given $X = x$, so that 'learning' the relation means to learn this conditional distribution (so as to enable accurate predictions for new, unseen data). Formally stated as Proposition 2.4, this connection highlights question (1) as essentially asking how to do supervised learning efficiently on sequential data, and it interprets identity (3) as a theoretically sound and practical signature-based answer to that question.

3 The Signature Representation of Sequential Data

While in Section 2 the relational data $\{(x,y)\} \subseteq \mathfrak{R}$ was not specified beyond belonging to some abstract vector space $\mathcal{X} \times \mathcal{Y}$, here we sharpen this assumption and operate on multidimensional sequential data, that is, statistically drawn data (8) with a temporal order to it. Let $d \in \mathbb{N}$ be fixed.

3.1 Sequential Data from Stochastic Processes

Given as an [in]finite tuple of vectors in $\mathbb{R}^d$, we call *sequential data* any ordered family

$$x \coloneqq (x_t \mid t \in I) \equiv (x_t)_{t\in I} \qquad \text{with} \qquad x_t \equiv (x_t^1, \ldots, x_t^d) \in \mathbb{R}^d \quad \text{and} \quad I \subset \mathbb{R}.$$

Denoting $Z \coloneqq \mathbb{R}^d$, we have $x \in Z^I$ and further assume this data to be ordered *continuously*, i.e.

$$x \in \left\{ y \in Z^I \mid I \ni t \mapsto y_t \ \text{ is continuous} \right\} =: \mathcal{C}(I; Z) \tag{12}$$

for significant technical convenience; note that there are many ways for discrete-time data to be embedded into $\mathcal{C}(I;Z)$, e.g. using piecewise-linear interpolation. Adding to (8) the continuity assumption (12), we understand the data-lift X in (8) to be continuous-time and Borel, that is

$$X : (\Omega, \mathscr{F}) \to (\mathcal{C}(I;Z), \|\cdot\|_\infty) \quad \text{is Borel-measurable.}$$

If I is compact, then usually $I = [0,1]$ wlog and we consider the Banach space

$$\left(\mathcal{C}_d \coloneqq \mathcal{C}([0,1];\mathbb{R}^d); \|\cdot\|_\infty\right) \ \text{ of all continuous paths } \ x \equiv (x_t) : [0,1] \to \mathbb{R}^d.$$

For simplicity, we will usually operate on the 'smooth core' $\mathcal{C}_d^1$ of absolutely continuous paths in $\mathcal{C}_d$,

$$\mathcal{C}_d^1 \coloneqq \left\{ x \in \mathcal{C}_d \;\middle|\; \exists!\, \dot{x} \in L^1([0,1];\mathbb{R}^d) \, : \, x_\cdot = x_0 + \int_0^\cdot \dot{x}_s \, \mathrm{d}s \right\}, \tag{13}$$

but, as usual, everything that follows admits canonical extensions to spaces of rougher paths as well. The subspace $\mathcal{C}_d^1$ of 'smooth' paths has the following convenient topological properties.

Lemma 3.1. *The space $\mathcal{C}_d^1$ is a Borel subset of $(\mathcal{C}_d, \|\cdot\|_\infty)$ and a separable Banach space wrt. the 1-variation norm $\|x\|_{1\text{-var}} = |x_0| + \|\dot{x}\|_{L^1}$, and the spaces $(\mathcal{C}_d^1, \|\cdot\|_\infty)$ and $(\mathcal{C}_d^1, \|\cdot\|_{1\text{-var}})$ have the same Borel σ-algebra.*

Proof. The first part of the lemma is well-known and the second part is proved in Appendix B.1. □

Set $\mathcal{X} \coloneqq \left(\mathcal{C}_{d_\mathcal{X}}^1, \|\cdot\|_{1\text{-var}}\right)$ and $\mathcal{Y} \coloneqq \left(\mathcal{C}_{d_\mathcal{Y}}^1, \|\cdot\|_{1\text{-var}}\right)$ for some $d_\mathcal{X}, d_\mathcal{Y} \in \mathbb{N}$. Given $(x,y) \in \mathcal{X} \times \mathcal{Y}$, let

$$\|(x,y)\|_{1\text{-var}} \coloneqq \sup_{\mathcal{D}} \sum_{(t_\nu)\in\mathcal{D}} |(x_{t_\nu}, y_{t_\nu}) - (x_{t_{\nu-1}}, y_{t_{\nu-1}})| \quad \text{and} \quad \|(x,y)\|_\infty \coloneqq \sup_{t\in[0,1]} |(x_t, y_t)|,$$

where the first supremum runs over the set $\mathcal{D}$ of all (finite) dissections (t_ν) of $[0,1]$.

The next lemma ensures that fusing two jointly distributed processes X and Y to a joint process (X,Y) ('input-output relation'; cf. Section 2) bears no measure-theoretic complications.

Lemma 3.2. *The space $\mathcal{Z} \coloneqq \left(\mathcal{X} \times \mathcal{Y}, \|\cdot\|_{1\text{-var}}\right)$ is Polish with Borel-σ-algebra $\mathcal{B}(\mathcal{Z}) \equiv \mathcal{B}(\mathcal{Z}, \|\cdot\|_{1\text{-var}}) = \mathcal{B}(\mathcal{Z}, \|\cdot\|_\infty)$. For $X : (\Omega, \mathscr{F}) \to \mathcal{X}$ and $Y : (\Omega, \mathscr{F}) \to \mathcal{Y}$, the joint process $Z \coloneqq (X,Y)$ is $\left(\mathscr{F}, \mathcal{B}(\mathcal{Z})\right)$-measurable iff X and Y are Borel-measurable.*

Proof. See Appendix B.2. □

3.2 The Signature Transform

After the general preparations of the previous subsection, we will now introduce the signature transform of a path [7, 28]. This transformation can be thought of as a faithful compression that sends a path to a hierarchically graded list of countably many coordinates that characterise the path. To define this transformation, we will first need to introduce some basic 'multiindex notation'.

Notation 3.3. Let $\boldsymbol{d} \in \mathbb{N}$. To properly index the announced coordinates of the signature, let

$$[\boldsymbol{d}]^* := \{\emptyset, \mathtt{1}, \mathtt{12}, \mathtt{21}, \mathtt{d11}, \mathtt{ddd1211d}, \ldots\}$$

be the free monoid over the alphabet $[\boldsymbol{d}] := \{\mathtt{1}, \mathtt{2}, \ldots, \mathtt{d}\}$, i.e. the monoid of all finite sequences of zero or more elements from $[\boldsymbol{d}]$, where $\emptyset$ denotes the empty word. The set $\mathbb{R}[\![\boldsymbol{d}]\!]$ $(\cong \mathbb{R}[\![\mathtt{x}_1, \ldots, \mathtt{x}_{\boldsymbol{d}}]\!])$ is the (free) algebra of all [multivariate] formal power series in the formal variables $\mathtt{1}\,(\cong \mathtt{x}_1), \ldots, \mathtt{m}\,(\cong \mathtt{x}_{\boldsymbol{d}})$. The element $\mathbf{1} := 1 \cdot \emptyset$ is the multiplicative unit in $\mathbb{R}[\![\boldsymbol{d}]\!]$. On a set-theoretic level

$$\mathbb{R}[\![\boldsymbol{d}]\!] = \{\boldsymbol{t} : [\boldsymbol{d}]^* \to \mathbb{R} \mid \boldsymbol{t} \text{ is a map}\} \equiv \left\{ \textstyle\sum_{w \in [m]^*} \boldsymbol{t}_w \cdot w \;\middle|\; \boldsymbol{t}_w \in \mathbb{R} \right\} \cong \textstyle\prod_{\nu=0}^{\infty} (\mathbb{R}^{\boldsymbol{d}})^{\otimes \nu}, \tag{14}$$

where the last identification is obtained by: identifying each word $\mathtt{i}_1 \cdots \mathtt{i}_{\boldsymbol{d}} \in [\boldsymbol{d}]^*$ with its associated elementary tensor $e_1 \otimes \cdots \otimes e_{\boldsymbol{d}} \in (\mathbb{R}^d)^{\otimes m}$, where $(e_i)_{i \in [\boldsymbol{d}]}$ is the standard basis in $\mathbb{R}^{\boldsymbol{d}}$.

On a geometrical note, let us for any path $z \equiv (z^1, \cdots, z^{\boldsymbol{d}}) \in \mathcal{C}_{\boldsymbol{d}}^{\text{1-var}}$ (of bounded 1-variation)[5] with components $z^1, \ldots, z^{\boldsymbol{d}} \in \mathcal{C}_1^{\text{1-var}}$ and any word $w \equiv \mathtt{i}_1 \mathtt{i}_2 \cdots \mathtt{i}_k \in [\boldsymbol{d}]^*$ $(k \in \mathbb{N})$ denote by

$$\mathrm{d}z^w \equiv \mathrm{d}z^{\mathtt{i}_1 \mathtt{i}_2 \cdots \mathtt{i}_k} := \mathrm{d}z_1^i \wedge \mathrm{d}z^{i_2} \wedge \cdots \wedge \mathrm{d}z^{i_k}$$

the w-indexed differential k-form defined via Lebesgue-Stieltjes differentials against the components (z^i) of z. Finally, let $\Delta_k := \{(t_\nu) \in [0,1]^k \mid 0 \leq t_1 \leq t_2 \leq \cdots \leq t_k \leq 1\}$ be the k-dimensional standard simplex, and define the length of a word w, denoted by $|w|$, as the number of its letters.◆

The following map is (essentially) a global coordinate chart that elucidates the space (13) of sequential data by injecting its elements to vectors in a Hilbert space which are easier to analyse.

Definition 3.4 (Signature)**.** The map $\mathfrak{sig} : \mathcal{C}_{\boldsymbol{d}}^{\text{1-var}} \to \mathbb{R}[\![\boldsymbol{d}]\!]$ that sends a path z to the formal series

$$\mathfrak{sig}(z) := \sum\nolimits_{w \in [\boldsymbol{d}]^*} \int_{\Delta_{|w|}} \mathrm{d}z^w \cdot w \;\cong\; \left(\int_{\Delta_{|w|}} \mathrm{d}z^w \;\middle|\; w \in \big[\boldsymbol{d}\big]^* \right) \tag{15}$$

is called the *signature*. More explicitly, its w-th coefficient for a path $z = (z_t^1, \cdots, z_t^d)_{t \in [0,1]}$ reads

$$\int_{\Delta_{|w|}} \mathrm{d}z^w \;\equiv\; \int_0^1 \int_0^{t_{k-1}} \int_0^{t_{k-2}} \cdots \int_0^{t_2} \int_0^{t_1} \mathrm{d}z_{t_0}^{i_1} \mathrm{d}z_{t_1}^{i_2} \cdots \mathrm{d}z_{t_{k-3}}^{i_{k-2}} \mathrm{d}z_{t_{k-2}}^{i_{k-1}} \mathrm{d}z_{t_{k-1}}^{i_k} \tag{16}$$

for the length-k $(= |w|)$ word $w = \mathtt{i}_1 \mathtt{i}_2 \cdots \mathtt{i}_{k-2} \mathtt{i}_{k-1} \mathtt{i}_k \in [\boldsymbol{d}]^*$ (so $\mathtt{i}_\nu \in \{\mathtt{1}, \ldots, \mathtt{d}\}$ for each $\nu \in [k]$). Given a path $x \in \mathcal{C}_d$, we denote its canonical monotone augmentation by

$$\bar{x} := (t, x_t)_{t \in [0,1]} \equiv \big(x_t^0, x_t^1, \cdots, x_t^d\big)_{t \in [0,1]} \in \mathcal{C}_{d+1} \quad \text{and set} \quad \bar{\iota} : \mathcal{C}_d \hookrightarrow \mathcal{C}_{d+1}, \quad x \mapsto \bar{\iota}(x) := \bar{x}, \tag{17}$$

for notational ease later on. The map (15) gives an embedding of paths into formal power series.

Theorem 3.5 ([16])**.** *The augmented signature map*

$$\underline{\mathfrak{sig}} := \mathfrak{sig} \circ \bar{\iota} \;:\; \mathcal{C}_d^1 \longrightarrow \mathbb{R}[\![d_0]\!] \equiv \mathbb{R}[\![\{0, 1, \ldots, d\}]\!] \quad \textit{is injective.} \tag{18}$$

Proof. Immediate by [16, Theorem 4] and the fact that, for any $x, y \in \mathcal{C}_d^1$, due to the strict monotonicity of their first component the augmented paths $\bar{x}$ and $\bar{y}$ are treelike equivalent iff $x = y$. □

Geometrically, the (augmented) signature (18) is a *global chart* for the Hilbert manifold $\mathcal{C}_d^1$ (Lemma 3.6). To exploit the transform (15) for our purposes, it will be convenient to endow its co-domain with some additional structure; more specifically, $\mathbb{R}[\![\boldsymbol{d}]\!]$ can be made into a Hilbert space.

[5] Here, we denote by $\mathcal{C}_{\boldsymbol{d}}^{\text{1-var}}$ the space of all the paths in $\mathcal{C}_{\boldsymbol{d}}$ that are of finite 1-variation.

3.2.1 Gradation and Inner Product (Of and Between Signature Vectors)

Any infinite $[\boldsymbol{d}]^*$-indexed tuple $\boldsymbol{a} \equiv (a_w)_{w\in[\boldsymbol{d}]^*} \subset \mathbb{R}$, such as (15), can be embedded into the algebra (14) via $\boldsymbol{a} = \sum_{w\in[\boldsymbol{d}]^*} \boldsymbol{t}_a(w)\cdot w \in \mathbb{R}[\![\boldsymbol{d}]\!]$ with $\boldsymbol{t}_a(\mathtt{i}_1\cdots\mathtt{i}_m) := a_{i_1\cdots i_m}$. Upon grouping the summands in (14) by their wordlength, we get

$$V \equiv \mathbb{R}[\![\boldsymbol{d}]\!] = \prod_{m=0}^{\infty} V_m \qquad \text{for} \qquad V_m := \bigoplus_{w\in[\boldsymbol{d}]^*\,:\,|w|=m} \mathbb{R}w\,, \tag{19}$$

where the *length* $|w| \in \mathbb{N}_0$ of a word $w \in [\boldsymbol{d}]^*$ is defined as the number of its letters. The gradation (19) comes with the canonical projections $\pi_m : V \to V_m$, $\pi_m(\boldsymbol{a}) := \boldsymbol{a}_m \equiv \sum_{|w|=m} a_w \cdot w$, for each $m \geq 0$, and we set $\pi_{[m]} \equiv \sum_{\nu=1}^m \pi_\nu \,:\, V \longrightarrow V_{[m]} := \bigoplus_{j=0}^m V_j$. Finally, we define the *inner product*

$$\langle\cdot\,,\cdot\rangle \,:\, V\times V \,\to\, \overline{\mathbb{R}}, \quad (\boldsymbol{a},\boldsymbol{b}) \,\mapsto\, \sum\nolimits_{w\in[\boldsymbol{d}]^*} \langle\boldsymbol{a},w\rangle\cdot\langle\boldsymbol{b},w\rangle =: \langle\boldsymbol{a},\boldsymbol{b}\rangle, \tag{20}$$

as the (infinite) bilinear extension of $\langle u,v\rangle := \delta_{uv}$, $u,v\in[\boldsymbol{d}]^*$. In other words, $[\boldsymbol{d}]^*$ is an ONS wrt. $\langle\cdot,\cdot\rangle$, and we note that $\langle\cdot,\cdot\rangle = \sum_{m\geq 0}\langle\pi_m(\cdot),\pi_m(\cdot)\rangle$ pointwise on $V\times V$.

Under convergence constraints, the above structure allows us to make V into a Hilbert space:

Lemma 3.6 (Hilbert Codomain of (15))**.** *For $(V,\langle\cdot,\cdot\rangle)$ the power series algebra from above, let*

$$\mathcal{H} \,:=\, \left\{\boldsymbol{t}\in V \;\middle|\; \|\boldsymbol{t}\| := \sqrt{\textstyle\sum_{m\geq 0}\|\pi_m(\boldsymbol{t})\|_m^2} < \infty\right\} \quad \textit{with} \quad \|\cdot\|_m := \sqrt{\langle\cdot,\cdot\rangle_m}\,, \tag{21}$$

where $\langle\cdot,\cdot\rangle_m := \langle\pi_m(\cdot),\pi_m(\cdot)\rangle$ for each $m\geq 0$. Then $(\mathcal{H},\langle\cdot,\cdot\rangle)$ is a separable Hilbert space with ONB $(w \mid w\in[\boldsymbol{d}]^*)$, which contains the image $\mathfrak{sig}(\mathcal{C}^1_{\boldsymbol{d}})$. Moreover, the maps

$$\mathfrak{sig} \,:\, \mathcal{C}^1_{\boldsymbol{d}} \to \mathcal{H} \qquad \textit{and} \qquad \underline{\mathfrak{sig}} \,:\, \mathcal{C}^1_{\boldsymbol{d}-1} \to \mathcal{H} \tag{22}$$

are both continuous wrt. the p-variation topology (on $\mathcal{C}^1_m$) for any $p\geq 1$.

Proof. The observations in this lemma are all well-known, see Appendix B.3. □

Writing $\xi_w \,:\, \mathcal{C}^1_{\tilde{d}} \ni x \mapsto \int_{\Delta_{|w|}} \mathrm{d}x^w \in \mathbb{R}$ for $w\in[\tilde{d}]^*$, where $\mathrm{d}x^{\mathtt{i}_1\cdots\mathtt{i}_m} := \mathrm{d}x^{i_1}\wedge\cdots\wedge\mathrm{d}x^{i_m}$, we have

$$\mathfrak{sig} \;=\; \sum\nolimits_{w\in[\tilde{d}]^*} \xi_w\cdot w$$

and this series $\|\cdot\|$-converges pointwise on $\mathcal{C}^1_{\tilde{d}}$ and uniformly over compact subsets of $\mathcal{C}^1_{\tilde{d}}$.

If desired, we can of course generalise (21) to a whole family of $\mathfrak{sig}$-containing Hilbert codomains:

Remark 3.7 (Alternative Hilbert Codomains)**.** Let $\gamma \equiv (\gamma_m)_{m\geq 0} > 0$ with $0 < \gamma_m \leq \lambda^m$ (all $m\in\mathbb{N}$) for some $\lambda > 0$. Given the gradation (19) of V together with the 'Euclidean identification' of its components $V_m \cong (V_1^{\otimes m},\langle\cdot,\cdot\rangle_m)$ seen above [right after (14)], with $V_1 \cong (\mathbb{R}^d,\langle\cdot,\cdot\rangle_2)$ and $\langle\cdot,\cdot\rangle_m \equiv \prod_{[m]}\langle\cdot,\cdot\rangle_2$ and $\|\cdot\|_m = \sqrt{\langle\cdot,\cdot\rangle_m^2}$, another natural Hilbert space structure on V is given by

$$\mathcal{H}^\gamma := \left\{\boldsymbol{t}\in V \;\middle|\; \|\boldsymbol{t}\|_\gamma := \sqrt{\textstyle\sum_{m\geq 0}\gamma_m\|\pi_m(\boldsymbol{t})\|_m^2} < \infty\right\} \tag{23}$$

together with the inner product $\langle\boldsymbol{s},\boldsymbol{t}\rangle_\gamma := \sum_{m\geq 0}\gamma_m\langle\pi_m(\boldsymbol{s}),\pi_m(\boldsymbol{t})\rangle_m$. It is clear that $(\mathcal{H}^\gamma,\langle\cdot,\cdot\rangle)$ is a Hilbert space (as the ℓ_2-direct sum of the Hilbert spaces $(V_m,\gamma_m\langle\cdot,\cdot\rangle_m)$, $m\geq 0$), see e.g. [12, Prop. I.6.2]), and we denote its topology by τ_γ. Clearly $\tau_{\tilde{\gamma}} \subseteq \tau_\gamma$ if $\tilde{\gamma}\leq\gamma$, as then $\|\cdot\|_{\tilde{\gamma}} \leq \|\cdot\|_\gamma$. ◆

3.3 The Signature as a 'Universal Nonlinearity'

A central property of the (augmented) signature transform (18) is that any bounded continuous function $f : (\mathcal{C}^1_d, \|\cdot\|_{1\text{-var}}) \to \mathbb{R}$ decomposes into a linear functional of the signature, or precisely: the diagram on the right commutes.

$$\begin{tikzcd}
\mathcal{C}^1_d \arrow[r, "\underline{\mathfrak{sig}}"] \arrow[dr, "f"'] & \mathbb{R}[\![d_0]\!] \arrow[r, dashed, "(\langle \tilde{f}_k, \cdot \rangle)"] & \ell_\infty \arrow[dl, "\vec{L}"] \\
 & \mathbb{R} &
\end{tikzcd}$$

Figure 1: *The signature lift as a 'universal nonlinearity': For every bounded continuous function f from $\mathcal{C}^1_d$ to $\mathbb{R}$, there exists a sequence $(\tilde{f}_k) \subset \mathbb{R}[d_0]$ such that $f = \lim_{k\to\infty}\langle \tilde{f}_k, \underline{\mathfrak{sig}}\rangle$ in the strict topology of Definition 3.8.*

(Here, $\vec{L} : \ell_\infty \to \mathbb{R}$ is the (unique, by Hahn-Banach) extension—from the [closed] subspace $\mathfrak{c}$ of convergent sequences to ℓ_∞—of the limit operator $\vec{L} : (\alpha_k) \mapsto \lim_{k\to\infty}\alpha_k$.)

Henceforth we employ the notational convention (cf. (18))

$$[d_0] \coloneqq \{0, 1, 2, \ldots, d\} \quad \text{and} \quad \bar{\iota}(x) =: (x^0_t, x^1_t, \cdots, x^d_t)_{t\in I},$$

thus $\langle \texttt{10230}, \underline{\mathfrak{sig}}(x)\rangle = \int_{\Delta_5} \mathrm{d}x^1 \wedge \mathrm{d}s \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3 \wedge \mathrm{d}t$ etc.

Let further $\mathbb{R}[d_0] \coloneqq \bigoplus_{m=0}^\infty V_m \subset \mathcal{H}$ be the subspace of $\mathbb{R}[\![d_0]\!]$ of all *polynomials* in the formal variables $[d_0]$.

The announced feature that nonlinear maps factor linearly through the signature is asymptotic, so we need to fix an appropriate topology for which convergence holds. Here, we follow [10] in using Giles' strict topology [15].

3.3.1 Some Topological Preparations

Let $\mathcal{X} \coloneqq (\mathcal{C}^1_d, \|\cdot\|_{1\text{-var}})$, write $C_b(\mathcal{X})$ for the set of all bounded continuous functions on $\mathcal{X}$, and write

$$B_0(\mathcal{X}) \coloneqq \{\psi : \mathcal{X} \to \mathbb{R} \text{ bounded} \mid \forall \varepsilon > 0 : \exists \mathcal{K} \subset \mathcal{X} \text{ compact } : \sup_{x\in\mathcal{X}\setminus\mathcal{K}} |\psi(x)| < \varepsilon\}$$

for the set of all bounded functions on $\mathcal{X}$ that *vanish at infinity.*

Definition 3.8 (Strict Topology [15])**.** The *strict topology* on $C_b(\mathcal{X})$, denoted by $\tau^{\mathcal{X}}_{\text{str}}$, is the topology induced by the family of seminorms

$$p_\psi(f) \coloneqq \sup_{x\in\mathcal{X}} \big|f(x)\psi(x)\big|, \qquad \psi \in B_0(\mathcal{X}).$$

Note that the strict topology is weaker than the uniform topology on $C_b(\mathcal{X})$ but stronger than the topology of compact (and thus also pointwise) convergence on $C_b(\mathcal{X})$, see [15, Theorem 2.4 (i)].

Lemma 3.9 ([15, Thm. 4.6]; Duality for $\tau^{\mathcal{X}}_{\text{str}}$)**.** *The topological dual of $\big(C_b(\mathcal{X}), \tau^{\mathcal{X}}_{\text{str}}\big)$ is the space of all finite signed regular Borel (fsrB) measures on $(\mathcal{X}, \|\cdot\|_\infty)$. More specifically: Given an fsrB measure μ on $\mathcal{B}(\mathcal{X})$, the linear functional $L_\mu : C_b(\mathcal{X}) \ni f \mapsto \int_{\mathcal{X}} f \,\mathrm{d}\mu \in \mathbb{R}$ is $\tau^{\mathcal{X}}_{\text{str}}$-continuous, and for each $\tau^{\mathcal{X}}_{\text{str}}$-continuous linear functional ϕ on $C_b(\mathcal{X})$ there exists a unique fsrB measure μ_ϕ such that $\phi = L_{\mu_\phi}$.*

Proof. This is simply [15, Theorem 4.6; see also Lemmas 4.2 and 4.5] combined with the identity $\sigma(\mathcal{X}, \|\cdot\|_{1\text{-var}}) = \sigma(\mathcal{X}, \|\cdot\|_\infty)$ from Lemma 3.1. □

As a further aid, we will build on the [10]-introduced concept of tensor normalisation to maintain the 'universal nonlinearity' property of the signature transform over unbounded path spaces, see Definition 3.14 below. To this end, we would like to make some preliminary observations.

Given $\lambda \in \mathbb{R}$, the λ-*dilation* is the map $\delta_\lambda : V \ni \boldsymbol{t} \mapsto \sum_{m=0}^\infty \lambda^m \pi_m(\boldsymbol{t}) \in V$ (thus, $\delta_1 = \mathrm{id}_V$).

Remark 3.10 (Signature Decay)**.** Let us recall from [29, Theorem 3.7 (case $p=1$)] that

$$\big\|\pi_m(\mathfrak{sig}(x))\big\|_m \;\le\; \|x\|^m_{1\text{-var}}/(m!\,\beta), \qquad \text{for each} \quad (x,m)\in\mathcal{X}\times\mathbb{N}_0 \tag{24}$$

(for some $[(x,m)$-independent$]$ constant $\beta>1$), i.e. the signature decays factorially. In particular,

$$\mathfrak{sig}(\mathcal{X})\;\subset\;\Big\{\boldsymbol{t}\in V \;\Big|\; \sum_{m\ge0}\|\pi_m(\boldsymbol{t})\|_m\lambda^m<\infty,\;\forall\,\lambda>0\Big\} =: \mathcal{H}_\downarrow \;\subseteq\; \mathcal{H},$$

and consequently $\delta_\lambda(\mathfrak{sig}(\mathcal{X}))\subset\mathcal{H}_\downarrow$ for each $\lambda>0$, since clearly $\delta_\lambda:\mathcal{H}_\downarrow\to\mathcal{H}_\downarrow$ for each $\lambda>0$. ◆

Lemma 3.11 ('Strong Continuity')**.** *Endow the subspace $\mathcal{H}_\downarrow$ with the locally convex topology $\tau_\downarrow$ which is induced by the family of norms*

$$|\!|\!|\cdot|\!|\!|_\lambda\,:\,\mathcal{H}_\downarrow\to\mathbb{R}, \quad \boldsymbol{t}\mapsto|\!|\!|\boldsymbol{t}|\!|\!|_\lambda := \sum_{m\ge0}\|\pi_m(\boldsymbol{t})\|\lambda^m, \qquad \lambda>0.$$

Then $(\mathcal{H}_\downarrow,\tau_\downarrow)$ is separable and metrizable Hausdorff and, for each $p\ge1$, the signature transform

$$\mathfrak{sig}\;:\;\big(\mathcal{X},\|\cdot\|_{p\text{-var}}\big)\;\to\;(\mathcal{H}_\downarrow,\tau_\downarrow) \quad \textit{is continuous.} \tag{25}$$

Proof. The topological qualities of $(\mathcal{H}_\downarrow,\tau_\downarrow)$ are due to [9, Corollary 2.4], while the continuity assertion about the signature is [9, Corollary 5.5]. □

Remark 3.12 (Comparison of Topologies)**.** Clearly, for any topological space $\mathcal{T}$, a map $\varphi:\mathcal{H}_\downarrow\to\mathcal{T}$ is $\tau_\downarrow$-continuous if φ is $|\!|\!|\cdot|\!|\!|_\lambda$-continuous for at least one $\lambda>0$. Moreover, note that $\mathcal{H}_\downarrow$ is a subspace of $\mathcal{H}^\gamma$, and that the above locally convex topology $\tau_\downarrow$ on $\mathcal{H}_\downarrow$ is *finer* than the (23)-induced subspace topology τ_γ on $\mathcal{H}_\downarrow$. Indeed: Since $\tau_\downarrow$ is metrizable (Lemma 3.11), the space $(\mathcal{H}_\downarrow,\tau_\downarrow)$ is sequential, whence $\tau_\gamma\subseteq\tau_\downarrow$ (iff $\mathrm{id}_{\mathcal{H}_\downarrow}:(\mathcal{H}_\downarrow,\tau_\downarrow)\to(\mathcal{H}_\downarrow,\tau_\gamma)$, $\boldsymbol{v}\mapsto\boldsymbol{v}$, is continuous) iff every $\tau_\downarrow$-convergent sequence in $\mathcal{H}_\downarrow$ is τ_γ-convergent. This clearly holds, however, since for every null-sequence $(\boldsymbol{v}_k)$ in $(\mathcal{H}_\downarrow,\tau_\downarrow)$ there is $k_0\in\mathbb{N}$ with $\sup_{k\ge k_0}|\!|\!|\boldsymbol{v}_k|\!|\!|_\lambda<1$ and hence, for each $k\ge k_0$,

$$\|\boldsymbol{v}_k\|^2_\gamma = \textstyle\sum_{m\ge0}\gamma_m\|\pi_m(\boldsymbol{v}_k)\|^2_m \;\le\; |\!|\!|\boldsymbol{v}_k|\!|\!|_\lambda \quad\longrightarrow\quad 0 \quad \text{as} \quad k\to\infty.$$

As one consequence of the inclusion $\tau_\gamma\subseteq\tau_\downarrow$, notice that (25) also holds for $\mathcal{H}_\downarrow$ replaced by $\mathcal{H}^\gamma$. ◆

Lemma 3.13. *If $\lambda_\cdot\,:\,(\mathcal{H}_\downarrow,\tau_\downarrow)\to\mathbb{R}_{>0}$ is a continuous positive scalar field, then the map*

$$\Lambda\;:\;(\mathcal{H}_\downarrow,\tau_\downarrow)\to(\mathcal{H}_\downarrow,\|\cdot\|), \quad \boldsymbol{t}\mapsto\delta_{\lambda_{\boldsymbol{t}}}\boldsymbol{t}, \quad \textit{is continuous.} \tag{26}$$

Consequently and for any fixed $p\ge1$, the Λ-scaled signature transform

$$\Lambda\circ\underline{\mathfrak{sig}}\;:\;(\mathcal{X},\|\cdot\|_{p\text{-var}})\;\to\;(\mathcal{H}_\downarrow,\|\cdot\|), \quad x\;\mapsto\;\delta_{\lambda_{\underline{\mathfrak{sig}}(x)}}(\underline{\mathfrak{sig}}(x)), \quad \textit{is continuous.} \tag{27}$$

Proof. Since the augmentation map $\bar\iota=(\theta,\mathrm{id}_{\mathcal{X}}):\mathcal{C}^1_d\to\mathcal{C}^1_{d_0}$ from (17), with $\theta(x):=(t)_{t\in[0,1]}$, is $(\|\cdot\|_{p\text{-var}},\|\cdot\|_{p\text{-var}})$-continuous, the continuity (27) follows immediately from (25) and assertion (26).

Let us now prove (26). Recall for this that $\Lambda(\boldsymbol{t})=\sum_{m\ge0}\lambda^m_{\boldsymbol{t}}\pi_m(\boldsymbol{t})$ by definition, and that, since $(\mathcal{H}_\downarrow,\tau_\downarrow)$ is first-countable by Lemma 3.11, the map Λ is continuous if it is sequentially continuous.

Let hence $\boldsymbol{t},(\boldsymbol{t}_k)\subset\mathcal{H}_\downarrow$ with $\lim_{k\to\infty}\boldsymbol{t}_k=\boldsymbol{t}$ in $\tau_\downarrow$, i.e. $\lim_{k\to\infty}|\!|\!|\boldsymbol{t}_k-\boldsymbol{t}|\!|\!|_{\tilde\lambda}=0$ for all $\tilde\lambda>0$. Then

$$\|\Lambda(\boldsymbol{t}_k)-\Lambda(\boldsymbol{t})\|^2\;\le\;2\sum_{m\ge0}\big(\|\lambda^m_{\boldsymbol{t}_k}\pi_m(\boldsymbol{t}_k)-\lambda^m_{\boldsymbol{t}_k}\pi_m(\boldsymbol{t})\|^2_m+\|\lambda^m_{\boldsymbol{t}_k}\pi_m(\boldsymbol{t})-\lambda^m_{\boldsymbol{t}}\pi_m(\boldsymbol{t})\|^2_m\big). \tag{28}$$

With $\lambda_{\cdot} : \boldsymbol{t} \mapsto \lambda_{\boldsymbol{t}}$ $\tau_{\downarrow}$-continuous, we have $\lim_{k\to\infty} |\lambda_{\boldsymbol{t}_k} - \lambda_{\boldsymbol{t}}| = 0$ and thus $c := \max\{\sup_k \lambda_{\boldsymbol{t}_k}, \lambda_{\boldsymbol{t}}\} < \infty$. As noted above, $\lim_{k\to\infty} \left\|\!\left|\boldsymbol{t}_k - \boldsymbol{t}\right|\!\right\|_c = 0$ and $\left\|\!\left|\boldsymbol{t}\right|\!\right\|_c < \infty$. The summands $\alpha_{m,k} := \|\lambda_{\boldsymbol{t}_k}^m \pi_m(\boldsymbol{t}_k - \boldsymbol{t})\|_m^2$ and $\beta_{m,k} := \|(\lambda_{\boldsymbol{t}_k}^m - \lambda_{\boldsymbol{t}}^m)\pi_m(\boldsymbol{t})\|_m^2$ on the right-hand side of (28) compare to

$$\alpha_{m,k} \;\le\; a_{m,k} := c^m \|\pi_m(\boldsymbol{q}_k - \boldsymbol{q})\|_m \quad \text{and} \quad \beta_{m,k} \;\le\; b_m := 4c^m \|\pi_m(\boldsymbol{q})\|_m, \tag{29}$$

for all $k \in \mathbb{N}$ and each $m \ge m_0$, for some sufficiently large $m_0 \in \mathbb{N}_0$. Hence and from (28) we find

$$\lim_{k\to\infty} \|\Lambda(\boldsymbol{t}_k) - \Lambda(\boldsymbol{t})\|^2 \;\le\; 2 \lim_{k\to\infty} \left\|\!\left|\boldsymbol{t}_k - \boldsymbol{t}\right|\!\right\|_c + \sum_{m\ge 0} \lim_{k\to\infty} \beta_{m,k} \;=\; 0\,, \tag{30}$$

where interchanging limit and summation for the second summand in (30) is permissible by dominated convergence, which in turn is applicable thanks to the $(\beta_{m,k})$-domination in (29) and the fact that $\sum_{m\ge 0} |b_m| = 4\left\|\!\left|\boldsymbol{q}\right|\!\right\|_c < \infty$. This proves (26), as desired. □

3.3.2 Universality of the Normalised Signature Transform

We can now recall the desired universality of flexibly scaled versions of the signature transform.

Definition 3.14 ([10])**.** We call *feature normalisation* (fN) any injective map of the form

$$\Lambda \;:\; \mathcal{H}_{\downarrow} \to \mathcal{H}_R := \{\boldsymbol{t} \in \mathcal{H} \mid \|\boldsymbol{t}\| \le R\}, \quad \boldsymbol{t} \;\mapsto\; \delta_{\lambda_{\boldsymbol{t}}}\,\boldsymbol{t}, \tag{31}$$

where $R > 0$ is a fixed constant and $\lambda : (\mathcal{H}_{\downarrow}, \tau_{\downarrow}) \ni \boldsymbol{t} \mapsto \lambda_{\boldsymbol{t}} \in \mathbb{R}_{>0}$ is continuous.

It's shown in [10, Sec. 3.2] that feature normalisations exist and can be conveniently constructed. As is also shown in [10], we can use (31) to 'squeeze' the signature's original coefficient functions (ξ_w) into $C_b(\mathcal{X})$ in such a way that most of the original signature's desirable structure is preserved.

Proposition 3.15 (Stone-Weierstrass for Bounded Signatures)**.** *Let $\Lambda = \delta_{\lambda_{(\cdot)}}$ be a feature normalisation, for $\lambda_{\cdot} : (\mathcal{H}_{\downarrow}, \tau_{\downarrow}) \to \mathbb{R}_{>0}$ continuous, and set $\underline{\lambda}_{(\cdot)} := \lambda_{\cdot} \circ \underline{\mathfrak{sig}}(\cdot)$. Then the family of signature coefficients*

$$\mathcal{A}_{\Lambda} \;:=\; \operatorname{span}_{\mathbb{R}}\Big\{ \underline{\xi}_w^{\lambda} \;:\; \mathcal{X} \ni x \;\mapsto\; \underline{\lambda}_x^{|w|} \int_{\Delta_{|w|}} \mathrm{d}\bar{x}^w \;\Big|\; w \in [d_0]^* \Big\}, \quad \textit{with} \quad \underline{\mathfrak{sig}}_{\Lambda} := \sum_{w \in [d_0]^*} \underline{\xi}_w^{\lambda} \cdot w, \tag{32}$$

is a point-separating and non-vanishing[6] *subalgebra of $C_b(\mathcal{X})$. Moreover, the algebra $\mathcal{A}_{\Lambda}$ is in fact dense in $\big(C_b(\mathcal{X}), \tau_{\mathrm{str}}^{\mathcal{X}}\big)$, which implies that*

$$\forall\, f \in C_b(\mathcal{X}) \;:\; \exists\, (\tilde{f}_k)_{k\in\mathbb{N}} \subset \mathbb{R}[d_0] \quad \textit{such that} \quad f = \lim_{k\to\infty} \big\langle \tilde{f}_k, \Lambda \circ \underline{\mathfrak{sig}} \big\rangle \quad \textit{in} \quad \tau_{\mathrm{str}}^{\mathcal{X}}. \tag{33}$$

If the domain $\mathcal{X}$ is in fact a bounded subset of $\mathcal{C}_d^1$, then all of the above holds for $\Lambda = \mathrm{id}_{\mathcal{H}_{\downarrow}}$.

Proof. This follows from [15, Theorem 3.1] via Theorem 3.5 and the fact that the span of all iterated integrals (16) is closed under multiplication, cf. also [10, Theorem 21] or see Appendix B.4 for a detailed proof. □

Their global boundedness and simultaneous universality, which come to fruition in Section 4.1 below, make the *normalised* signatures $\underline{\mathfrak{sig}}_{\Lambda} = \Lambda \circ \underline{\mathfrak{sig}}$ from (32) (also called 'robust signatures' in [10]) our transformations of choice to derive the desired representations (2) and (3).

[6] A family $\mathcal{A} \subseteq C(\mathcal{X})$ will be called *non-vanishing* if: $\forall\, x \in \mathcal{X}$ there exists $\varphi \in \mathcal{A}$ such that $\varphi(x) \neq 0$.

4 Computing Conditional Expectations via Signatures

Let us return to the stochastic setting (10) where we consider a pair (X, Y) of $(\mathscr{X} \times \mathscr{Y})$-valued random variables to make the learning of relations accessible, via lifting, to a rigorous probabilistic description. Suppose for the following that the data of interest is sequential as in Section 3.1, and specifically that

$$\mathscr{X} \coloneqq \big(\mathcal{C}^1_{d_{\mathscr{X}}}, \|\cdot\|_{1\text{-var}}\big) \quad \text{and} \quad \mathscr{Y} \coloneqq \big(\mathcal{C}^1_{d_{\mathscr{Y}}}, \|\cdot\|_{1\text{-var}}\big) \qquad \text{for some} \quad d_{\mathscr{X}}, d_{\mathscr{Y}} \in \mathbb{N}.$$

In this case, the random variables, supported on the same (complete) probability space $(\Omega, \mathscr{F}, \mathbb{P})$,

$$X \,:\, \Omega \to \mathscr{X} \quad \text{and} \quad Y \,:\, \Omega \to \mathscr{Y}, \tag{34}$$

are continuous-time *stochastic processes*. Remark A.1 collects some useful technical details on (34).

The random variable X in (34) induces a sub-σ-algebra $\Sigma_X \coloneqq \sigma(X)$ of $\mathscr{F}$ that serves us as the 'informational basis' for an X-informed [L^2-] best-approximation Y_X of the process Y. Specifically,

$$Y_X \;=\; \mathbb{E}[Y \,|\, X] \,:\, \Omega \to \mathscr{Y}, \tag{35}$$

i.e., the X-based [L^2-]optimal proxy Y_X of Y is given as the conditional expectation of Y wrt. Σ_X.[7]

Now an approximation Y_X of the full process Y is generally difficult to come by, especially because of the 'analytical intractability' of the path space $\mathscr{Y}$, but often also not directly required for applications. Instead, one is usually more interested in the *conditional distribution*

$$\mathcal{B}(\mathscr{Y}) \ni A \;\longmapsto\; \mathbb{P}(Y \in A \,|\, X) \coloneqq \mathbb{E}[\mathbb{1}_A(Y) \,|\, \Sigma_X] \tag{36}$$

of Y given X, and in associated derived statistics of Y, as we have seen in Sections 1 and 2.

Delivering on the initial announcement (2), this section presents a new approach to first compute (35) efficiently and with controllable precision (Theorem 4.5 and Corollary 4.6). According representations for (36) then follow in Section 5.1 (Proposition 5.1 and Corollary 5.2).

4.1 The Conditional Expected Signature of Y given X

The processes X and Y, given as the Banach-space-valued random variables (34) per default, can be analysed more profitably in the 'time-global' coordinates (16) provided by the Hilbert-charts (22).

So we transition from the default description (34) to the Hilbert setting (22) and abbreviate

$$\mathbb{X} \coloneqq \underline{\mathfrak{sig}}(X) \qquad \text{and} \qquad \mathbb{Y} \coloneqq \underline{\mathfrak{sig}}(Y). \tag{37}$$

We know from Lemma 3.6 that $\mathbb{X}$ and $\mathbb{Y}$ are Hilbert-space-valued random variables[8], specifically:

$$\mathbb{X} \in \mathcal{L}(\mathbb{P}, \mathcal{H}_{\mathscr{X}}) \coloneqq \mathcal{L}(\Omega, \mathscr{F}, \mathbb{P}\,;\, \mathcal{H}_{d_{\mathscr{X}}}) \qquad \text{and} \qquad \mathbb{Y} \in \mathcal{L}(\mathbb{P}, \mathcal{H}_{\mathscr{Y}}),$$

where $\mathcal{H}_d$ ($d \in \mathbb{N}$) is the Hilbert subset of $\mathbb{R}[\![d]\!]$ of square-summable power series as defined in (21).

Differentiating degrees $1 \leq p < \infty$ of integrability, we also introduce the spaces and norms

$$L^p(\mathbb{P}, \mathcal{H}) \coloneqq \big\{\mathbb{Z} \in \mathcal{L}(\mathbb{P}, \mathcal{H}) \;\big|\; \mathbb{E}\|\mathbb{Z}\|^p < \infty\big\} \quad \text{and} \quad \big\|\mathbb{Z}\big\|_{L^p(\mathcal{H})} \coloneqq \left(\int_\Omega \big\|\mathbb{Z}(\omega)\big\|^p \mathrm{d}\mathbb{P}\right)^{1/p} = \mathbb{E}\big[\|\mathbb{Z}\|^p\big]^{\frac{1}{p}}. \tag{38}$$

[7] Recall that the conditional expectation (35) exists if Y is Bochner-integrable [which we don't need to assume for our actual purposes], and it is unique up to $\mathbb{P}$-almost sure equality; see e.g. [36, Theorem II.2.1].

[8] Here as throughout, any random variable is understood to be Borel: $\mathbb{Z} \in \mathcal{L}(\mathbb{P}, \mathcal{H}) :\Leftrightarrow \mathbb{Z}$ is $\big(\mathscr{F}, \sigma(\|\cdot\|_{\mathcal{H}})\big)$-measurable ($\Leftrightarrow \mathbb{Z}$ is Bochner-measurable, since the $\mathcal{H}$ are separable [Lemma 3.6]; see e.g. [42, Proposition 1.8]).

Recall that $(L^p(\mathbb{P},\mathcal{H}),\|\cdot\|_{L^p(\mathcal{H})})$ is Banach for $1 \leq p < \infty$ and Hilbert for $p = 2$ with inner product

$$\big\langle \mathbb{Z}_1, \mathbb{Z}_2 \big\rangle_{L^2(\mathcal{H})} \;\coloneqq\; \int_\Omega \langle \mathbb{Z}_1(\omega), \mathbb{Z}_2(\omega)\rangle \,\mathrm{d}\mathbb{P} \;=\; \mathbb{E}\big[\langle \mathbb{Z}_1, \mathbb{Z}_2\rangle\big]. \tag{39}$$

To avoid any integrability concerns on side of the coordinate representation $\mathbb{Y}$ of Y, consider

$$\mathbb{Y}^\Lambda \;\coloneqq\; \Lambda(\mathbb{Y}) \qquad \text{for} \quad \Lambda : \mathcal{H}_\mathcal{Y}^\downarrow \to \mathcal{H}_\mathcal{Y} \quad \text{some fixed feature normalisation;} \tag{40}$$

$$\mathbb{X}^\Xi \;\coloneqq\; \Xi(\mathbb{X}) \qquad \text{for} \quad \Xi : \mathcal{H}_\mathcal{X}^\downarrow \to \mathcal{H}_\mathcal{X} \quad \text{some fixed feature normalisation.} \tag{41}$$

Let us make a few simple but far-reaching observations.

Lemma 4.1. *We have that $\mathbb{Y}^\Lambda \in L^p(\mathbb{P},\mathcal{H}_\mathcal{Y})$ for each $p \geq 1$.*

Proof. The fact that $\mathbb{Y}^\Lambda \in \mathcal{L}(\mathbb{P},\mathcal{H}_\mathcal{Y})$ is due to (27) and the fact that in our setting, Bochner- and Borel-measurability coincide thanks to the separability of $\mathcal{H}_\mathcal{Y}$ (e.g. [42, Prop. 1.8]). The unrestricted integrability of $\mathbb{Y}^\Lambda$ is clear since Λ is bounded by definition (31) of an fN. □

Lemma 4.2. *Adopting the setting and notation of Proposition 3.15, we have that*

$$\Sigma_X \;=\; \sigma\big(\underline{\mathfrak{sig}}_\Lambda(X)\big) \;=\; \sigma\Big(\underline{\xi}^\lambda_w(X) \;\Big|\; w \in [d_0]^*\Big). \tag{42}$$

Proof. Let us abbreviate $\varphi \coloneqq \underline{\mathfrak{sig}}_\Lambda$ and $\phi_w \coloneqq \underline{\xi}^\lambda_w(X)$ for each $w \in [d_0]^*$. We first prove that $\sigma(X) = \sigma(\varphi(X))$: The inclusion $\sigma(\varphi(X)) \subseteq \sigma(X)$ is immediate since $\varphi : \mathcal{X} \to \mathcal{H}_{d_0}$ is continuous (and hence Borel-measurable), see Lemma 3.13. For the converse, note that since φ is also an injection, we have that $A_\varphi \coloneqq \varphi(A) \in \mathcal{B}(\mathcal{H}_{d_0}) \cap \varphi(\mathcal{X})$ for any fixed $\|\cdot\|_{\text{1-var}}$-open set $A \subseteq \mathcal{X}$. Indeed, the set A_φ is analytic (as the continuous image of a Borel subset of a Polish space) and so is its complement $A_\varphi^c \equiv \varphi(\mathcal{X}) \setminus A_\varphi = \varphi(A^c)$, where the last identity holds since φ is injective; this implies that $A_\varphi \in \mathcal{B}(\mathcal{H}_{d_0} \cap \varphi(\mathcal{X}))$ by a theorem of Souslin [23, Corollary 3.1 (p. 486)]. Consequently, $X^{-1}(A) = (\varphi(X))^{-1}(A_\varphi) \in \sigma(\varphi(X))$ and hence $\sigma(X) \subseteq \sigma(\varphi(X))$, as desired. Let us next prove the second identity in (42).

The inclusion $\sigma(\phi_w \mid w \in [d_0]^*) \subseteq \sigma(\varphi(X))$ is immediate since $\phi_w = \langle w, \varphi(X)\rangle$ for each $w \in [d_0]^*$ (cf. (20)) and each $\langle w, \cdot\rangle : \mathcal{H}_{d_0} \to \mathbb{R}$ is continuous. The converse inclusion holds as, pointwise on Ω,

$$\|\varphi(X) - \psi_n\| \;\leq\; \sum_{m=n+1}^{\infty} \|\pi_m(\varphi(X))\|_m \;\overset{m\to\infty}{\longrightarrow}\; 0, \qquad \text{for} \quad \psi_n \coloneqq \textstyle\sum_{|w|\leq n} \phi_w \cdot w$$

(cf. (32) and (27)), i.e. $\varphi(X)$ is the pointwise limit of $\sigma(\phi_w |\, w)$-measurable functions and thus $\sigma(\phi_w |\, w)$-measurable itself. □

Setting $(d,\tilde{d}) \coloneqq \big((d_\mathcal{X})_0, (d_\mathcal{Y})_0\big)$ and for any fixed fN Ξ on $\mathcal{H}_\mathcal{X}$, let us introduce the *coefficient space*

$$\mathfrak{L}^2_X \;\coloneqq\; \big\{\alpha : \big[\tilde{d}\big]^* \to \mathbb{R}[d],\, w \mapsto \alpha_w \;\big|\; \|\alpha\|^2_\mathfrak{L} \coloneqq \textstyle\sum_{w\in[\tilde{d}]^*} \mathbb{E}\big[\langle \alpha_w,\, \mathbb{X}^\Xi\rangle^2\big] < \infty\big\}$$

(cf. (41) for notation), as well as the space

$$L^2_X(\mathcal{H}_\mathcal{Y}) \coloneqq L^2(\mathbb{P},\Sigma_X;\mathcal{H}_\mathcal{Y}) \;\equiv\; \{\mathbb{Z} \in L^2(\mathbb{P},\mathcal{H}_\mathcal{Y}) \mid \mathbb{Z} \sim_{L^2} \tilde{\mathbb{Z}} \,:\, \tilde{\mathbb{Z}}^{-1}(\mathcal{B}(\mathcal{H}_\mathcal{Y})) \subseteq \Sigma_X\} \tag{43}$$

of all Σ_X-measurable mean-square integrable $\mathcal{H}_\mathcal{Y}$-valued random variables.

The following result provides us with a large family of easily adjustable model functions that will allow us to compute the conditional expectation $\mathbb{E}[\mathbb{Y}^\Lambda \,|\, X]$ as the unique solution of a convex and feasibly implementable 'least-squares type' optimisation problem.

Proposition 4.3. *For Ξ as above, consider the family of $\mathfrak{L}^2_X$-parametrised 'simple' functions*

$$\Psi \coloneqq \{\psi_\alpha : \mathcal{X} \to \mathcal{H}_{d_\mathcal{Y}} \cup \{\infty\} \mid \alpha \equiv (\alpha_w) \in \mathfrak{L}^2_X\} \quad \textit{given by} \quad \psi_\alpha \coloneqq \sum_{w\in[\tilde{d}]^*} \langle \alpha_w, \underline{\mathfrak{sig}}_{\Xi}(\,\cdot\,)\rangle \cdot w. \tag{44}$$

Then, the family of 'simple' random variables

$$\Psi(X) \coloneqq \{\psi(X) \mid \psi \in \Psi\} \quad \textit{is an } \|\cdot\|_{L^2(\mathcal{H}_\mathcal{Y})}\textit{-dense subset of } L^2_X(\mathcal{H}_\mathcal{Y}). \tag{45}$$

Proof. We first show that $\Psi(X) \subseteq L^2_X(H_\mathcal{Y})$. Let for this $\alpha \in \mathfrak{L}^2_X$ be fixed. By monotone convergence,

$$\|\psi_\alpha(X)\|^2_{L^2(\mathcal{H}_\mathcal{Y})} = \textstyle\sum_{w\in[\tilde{d}]^*}\mathbb{E}\big[|\langle \alpha_w, \underline{\mathfrak{sig}}_{\Xi}(X)\rangle|^2\big] = \|\alpha\|^2_{\mathfrak{L}} < \infty.$$

In particular, $\|\psi_\alpha(X)\| < \infty$ almost surely, whence the $(\Sigma_X, \mathcal{B}(\mathcal{H}_\mathcal{Y}))$-measurability of $\psi_\alpha(X)$ follows via Pettis measurability theorem (e.g. [42, Theorem 1.11], applicable by Remark A.1 (iv)) from the facts that the compositions $\langle w, \psi_\alpha(X)\rangle = \langle \alpha_w, \underline{\mathfrak{sig}}_{\Xi}(X)\rangle$ are each Σ_X-measurable by Lemmas 3.13 & 4.2 and $(\langle w, \cdot\rangle \mid w \in [\tilde{d}])$ is a (Schauder) basis of the [topological] dual of $\mathcal{H}_{\tilde{d}}$. (In our setting the notions Bochner- and Borel-measurability coincide thanks to the separability of $\mathcal{H}_{\tilde{d}}$, see e.g. [42, Prop. 1.8].)

Next we show the density assertion (44). Our proof uses the following classical result:

> **Lemma 4.4** (**Functional Monotone Class**)**.** *Suppose that H is a vector space of bounded real-valued functions on a measurable space $\mathscr{X}$ such that H contains the constants and is closed under bounded monotone convergence (that is, for any increasing sequence $(\varphi_k) \subset H$ of positive, uniformly bounded functions, the (pointwise) limit $\varphi \coloneqq \lim_{k\to\infty}\varphi_k$ lies in H). Let $\mathfrak{C}$ be a subset of H which is closed under pointwise multiplication, then H contains all $\sigma(\mathfrak{C})$-measurable bounded functions.*
>
> *Proof of Lemma 4.4.* See for instance [18, Theorem A.1]. □

For arbitrary $\mathbb{Z} \in L^2_X(\mathcal{H}_\mathcal{Y}) \eqqcolon G$ and $w \in [\tilde{d}]^*$ fixed, note that $\mathbb{Z}_w \coloneqq \langle w, \mathbb{Z}\rangle \in L^2(\Omega, \Sigma_X, \mathbb{P}) \eqqcolon G_1$. (Indeed, $\mathbb{E}\big[|\mathbb{Z}_w|^2\big] \leq \sum_{w\in[\tilde{d}]^*}\mathbb{E}\big[|\mathbb{Z}_w|^2\big] = \mathbb{E}\big[\sum_{w\in[\tilde{d}]^*}\mathbb{Z}_w^2\big] = \mathbb{E}\big[\|\mathbb{Z}\|^2\big] = \|\mathbb{Z}\|^2_{L^2(\mathcal{H}_\mathcal{Y})} < \infty$.) Now let

$$H \coloneqq \overline{\mathcal{A}_\Xi(X)}^{L^2} \cap L^\infty \quad \text{and} \quad \mathfrak{C} \coloneqq \mathcal{A}_\Xi(X), \quad \text{where} \quad \mathcal{A}_\Xi(X) \coloneqq \{\xi(X) \mid \xi \in \mathcal{A}_\Xi\} \quad (\text{cf. } (32)). \tag{46}$$

In other words, $\mathfrak{C}$ is the vector space of images of X under the maps from $\mathcal{A}_\Xi$, where $\mathcal{A}_\Xi$ the space of all Ξ-scaled signature polynomials as defined by $(32)|_{\Lambda=\Xi}$, and H is the set of all bounded Σ_X-measurable random variables which are in the G_1-closure of $\mathfrak{C}$. From Proposition 3.15 we know that $\mathcal{A}_\Xi$ is a subalgebra of $C_b(\mathcal{X})$, and consequently $\mathfrak{C}$ is a subset of H which is closed under pointwise multiplication. Next, let us show that H satisfies the hypotheses of Lemma 4.4: First, it is clear that H is a vector space (as the intersection of two vector spaces) which also contains the constants since $\mathfrak{C}$ contains the constants. To check for the appropriate closedness of H, note that for any monotone sequence $(\mathbb{z}_k) \subset H$ such that $\sup_k|\mathbb{z}_k| \leq C$ for some $C > 0$ and $\mathbb{z} \coloneqq \lim_{k\to\infty}\mathbb{z}_k$ pointwise, we have $\mathbb{z}_k \to \mathbb{z}$ in L^2 by dominated convergence, implying $\mathbb{z} \in H$ as required.

The above pair $(H, \mathfrak{C})$ thus qualifies for the application of Lemma 4.4, which yields that H contains all bounded $\sigma(\mathfrak{C})$-measurable functions. But since $\mathfrak{C} = \operatorname{span}_\mathbb{R}\{\underline{\xi}^{\Xi}_w(X) \mid w \in [\tilde{d}]\}$ and hence $\sigma(\mathfrak{C}) = \sigma(\underline{\xi}^{\Xi}_w(X) \mid w \in [\tilde{d}])$, Lemma 4.2 implies $\sigma(\mathfrak{C}) = \Sigma_X$, which shows that in fact we have proved

$$G_1 \cap L^\infty \subseteq H \subseteq \overline{\mathfrak{C}}^{L^2}. \tag{47}$$

Before using this observation to prove (45), note that for the above w-coordinate $\mathbb{Z}_w$ we have

$$\mathbb{Z}_w = L^2\text{-}\lim_{n\to\infty}\mathbb{Z}^{\langle n\rangle}_w \quad \text{for the truncations} \quad \mathbb{Z}^{\langle n\rangle}_w \coloneqq \max\big(-n, \min(\mathbb{Z}_w, n)\big) \in G_1 \cap L^\infty$$

(the above L^2-convergence holds by dominated convergence). But since $(\mathbb{Z}_w^{\langle n\rangle}) \subset \overline{\mathfrak{C}}^{L^2}$ by (47), we find

$$\mathbb{Z}_w \,\in\, \overline{\mathfrak{C}}^{L^2}, \quad \text{that is:} \quad \mathbb{Z}_w \,=\, L^2\text{-}\lim_{j\to\infty} \langle \alpha_{w,j},\, \underline{\mathfrak{sig}}_{\Xi}(X)\rangle \quad \text{for some} \;\; (\alpha_{w,j})_j \subset \mathbb{R}[d]. \tag{48}$$

Since (48) holds for all $w \in [\tilde{d}]^*$, the desired conclusion (45) is now within very close reach:

Fix any $\varepsilon > 0$. Abbreviating $\varphi_p := \langle p, \underline{\mathfrak{sig}}_{\Xi}(X)\rangle$ for $p \in \mathbb{R}[d]$, choose some

$$\alpha_w^\star \,\in\, \mathbb{R}[d] \quad \text{such that} \quad \|\mathbb{Z}_w - \varphi_{\alpha_w^\star}\|_{G_1}^2 \,\leq\, \varepsilon^2(2\tilde{d})^{-|w|}/2 \quad \big(w \in [\tilde{d}]^*\big).$$

Then for the coefficient vector $\alpha^\star := (\alpha_w^\star)_{w\in[\tilde{d}]^*}$ we get that

$$\|\alpha^\star\|_{\mathfrak{L}}^2 \,=\! \sum_{w\in[\tilde{d}]^*} \mathbb{E}\big[\varphi_{\alpha_w^\star}^2\big] \,\leq\, 2\sum_{m=0}^{\infty}\sum_{|w|=m} \|\mathbb{Z}_w - \varphi_{\alpha_w^\star}\|_{G_1}^2 + \|\mathbb{Z}_w\|_{G_1}^2 \,\leq\, \varepsilon^2 \sum_{m=0}^{\infty} 2^{-m} \,+\, 2\|\mathbb{Z}\|_G^2 \,<\, \infty, \tag{49}$$

where the penultimate inequality is due to there being $\sharp\{w \in [\tilde{d}]^* \mid |w| = m\} = \tilde{d}^m$ many words of length m in the index-monoid $[\tilde{d}]^*$. Hence $\alpha^\star \in \mathfrak{L}_X^2$ and thus $\psi_{\alpha^\star} \in \Psi(X)$, and from (49) we read off

$$\|\mathbb{Z} - \psi_{\alpha^\star}\|_{L^2(\mathcal{H}_\mathcal{Y})}^2 \,=\, \sum_{m=0}^{\infty}\sum_{|w|=m} \|\mathbb{Z}_w - \varphi_{\alpha_w^\star}\|_{G_1}^2 \,\leq\, \varepsilon^2.$$

Since both $\mathbb{Z} \in L_X^2(\mathcal{H}_\mathcal{Y})$ and $\varepsilon > 0$ were arbitrary, the claim (45) is established. □

Combining Proposition 4.3 with the classical perspective on conditional expectation as an L^2-projection then yields the following *variational characterisation* of $\mathbb{E}[\mathbb{Y}^\Lambda \,|\, X]$.

Theorem 4.5. *Adopting the setting and notation of Proposition 4.3, we have that*

$$\mathbb{E}\big[\mathbb{Y}^\Lambda \,\big|\, X\big] \,=\, \lim_{k\to\infty} \psi_{\alpha_k}(X) \quad \textit{in} \;\; L_X^2(\mathcal{H}_\mathcal{Y}) \tag{50}$$

for any minimizing sequence $(\alpha_k) \subset \mathfrak{L}_X^2$ of the (convex) infinite linear least squares problem

$$\inf_{\alpha\in\mathfrak{L}_X^2} \mathbb{E}\big[\|\mathbb{Y}^\Lambda - \psi_\alpha(X)\|^2\big]. \tag{51}$$

The convergence (50) holds $\mathbb{P}$-almost surely if (α_k) is such that $\sum_{k=0}^{\infty}(\Phi(\psi_{\alpha_k}(X)) - \gamma)^{1/2} < \infty$, where $\Phi(\mathbb{Z}) := \mathbb{E}[\|\mathbb{Y}^\Lambda - \mathbb{Z}\|^2]$ and $\gamma := \inf_{\alpha\in\mathfrak{L}_X^2} \Phi(\psi_\alpha(X))$.

Proof. We begin with the well-known observation that the space $L_X^2(\mathcal{H}_\mathcal{Y})$ from (43) is a closed linear subspace of $\big(L^2(\mathbb{P}, \mathcal{H}_\mathcal{Y}), \|\cdot\|_{L^2(\mathcal{H}_\mathcal{Y})}\big)$, which entails the $\langle\cdot,\cdot\rangle_{L^2(\mathcal{H}_\mathcal{Y})}$-orthogonal decomposition

$$L^2(\mathbb{P}, \mathcal{H}_\mathcal{Y}) \,=\, L_X^2(\mathcal{H}_\mathcal{Y}) \,\oplus\, L_X^2(\mathcal{H}_\mathcal{Y})^\perp.$$

(The closedness of $L_X^2(\mathcal{H}_\mathcal{Y})$ follows from the well-known fact (which persists for Hilbert-valued random variables [42, Proposition 2.11]) that L^2-convergence implies almost sure convergence on a subsequence.) Denoting $E := L^2(\mathbb{P}, \mathcal{H}_\mathcal{Y})$ and $G := L_X^2(\mathcal{H}_\mathcal{Y})$ for brevity, we then adopt (from the scalar-valued setting, e.g. [42, Section 11.1]) the classical perspective that the (vector-valued) conditional expectation $\mathbb{Y}_X^\Lambda := \mathbb{E}\big[\mathbb{Y}^\Lambda \,\big|\, X\big] \in G$ is the orthogonal projection of $\mathbb{Y}^\Lambda$ onto G along $G^\perp$, in symbols:

$$\mathbb{Y}_X^\Lambda \,=\, \mathrm{P}_G\mathbb{Y}^\Lambda \quad \text{for the orthogonal projector} \;\; \mathrm{P}_G : E \to E \;\; \text{on} \;\; G = \mathrm{im}(\mathrm{P}_G). \tag{52}$$

To see that this perspective (52) is true in the present vector-valued setting (38) & (39), denote $\mathbb{Y}_G := \mathrm{P}_G\mathbb{Y}^\Lambda$ and notice that then $\Delta := \mathbb{Y}^\Lambda - \mathbb{Y}_G \in G^\perp$, that is $\langle \Delta, \chi\rangle_{L^2(\mathcal{H}_\mathcal{Y})} = 0$ for all $\chi \in G$. In particular,

$$0 \;=\; \langle \Delta, w\mathbb{1}_A\rangle_{L^2(\mathcal{H}_\mathcal{Y})} = \langle \Delta\mathbb{1}_A, w\rangle_{L^2(\mathcal{H}_\mathcal{Y})} \;=\; \langle \int_A \mathbb{Y}^\Lambda \,\mathrm{d}\mathbb{P}, w\rangle - \langle \int_A \mathbb{Y}_G \,\mathrm{d}\mathbb{P}, w\rangle \quad \big(A \in \Sigma_X,\;\; w \in [\tilde{d}]^*\big),$$

where the last identity is due to Bochner integrals commuting with bounded linear functionals, cf. [42, Sec. 1.3.1]; note that each element of E is Bochner-integrable by definition (38) and [42, Prop. 1.16]. Since $[\tilde{d}]^*$ is an orthonormal basis of $\mathcal{H}_\mathcal{Y}$, the above implies that: $\int_A \mathbb{Y}^\Lambda \,\mathrm{d}\mathbb{P} = \int_A \mathbb{Y}_G \,\mathrm{d}\mathbb{P}$, for all $A \in \Sigma_X$. But the latter property is characteristic also of the vector-valued conditional expectation $\mathbb{E}[\mathbb{Y}^\Lambda \,|\, \Sigma_X]$, see e.g. [42, Theorem 11.10], which implies that $\mathbb{Y}_G = \mathbb{Y}^\Lambda_X$ as claimed in (52).

The above characterisation (52) of $\mathbb{Y}^\Lambda_X$ as the orthogonal projection of $\mathbb{Y}^\Lambda$ onto G implies that

$$\big\|\mathbb{Y}^\Lambda - \mathbb{Y}^\Lambda_X\big\|_{L^2(\mathcal{H}_\mathcal{Y})} \leq\; \big\|\mathbb{Y}^\Lambda - \mathbb{Z}\big\|_{L^2(\mathcal{H}_\mathcal{Y})} \quad \text{for all}\;\; \mathbb{Z} \in G. \tag{53}$$

Moreover, the Hilbert projection theorem guarantees that the arg min in (53) is unique, so that in fact

$$\mathbb{Y}^\Lambda_X \;=\; \operatorname*{arg\,min}_{\mathbb{Z}\in G} \mathbb{E}\big\|\mathbb{Y}^\Lambda - \mathbb{Z}\big\|^2. \tag{54}$$

Now in order to make the variational identity (54) more operational, recall from Prop. 4.3 that

$$\text{the set}\;\; \mathfrak{G} := \Psi(X) = \big\{\psi_\alpha(X) \;\big|\; \alpha \in \mathfrak{L}^2_X\big\} \;\;\text{from}\;\; (45) \;\;\text{is}\;\; \|\cdot\|_{L^2(\mathcal{H}_\mathcal{Y})}\text{-dense} \;\;\text{in}\;\; G. \tag{55}$$

Let us observe how (55) and (54) imply (50): Abbreviating $\Phi(\mathbb{Z}) := \|\mathbb{Y}^\Lambda - \mathbb{Z}\|^2_{L^2(\mathcal{H}_\mathcal{Y})}$, which defines a function $\Phi : G \to \mathbb{R}_+$ that is clearly $\|\cdot\|_{L^2(\mathcal{H}_\mathcal{Y})}$-continuous and strictly convex, we find that

$$\Phi(\mathbb{Y}^\Lambda_X) \overset{(54)}{=} \inf_{\mathbb{Z}\in G} \Phi(\mathbb{Z}) \overset{(55)}{=} \inf_{\mathbb{Z}\in\mathfrak{G}} \Phi(\mathbb{Z}) \;=\; \inf_{\alpha\in\mathfrak{L}^2_X} \Phi(\psi_\alpha(X)) \;=:\; \gamma.$$

Hence for any minimizing sequence of (51), i.e. any sequence (α_k) in $\mathfrak{L}^2_X$ with $\lim_{k\to\infty} \Phi(\psi_{\alpha_k}(X)) = \gamma$, the functions $(\mathbb{Z}_k) := (\psi_{\alpha_k}(X)) \subset G$ are a minimizing sequence for $\inf_{\mathbb{Z}\in G} \Phi(\mathbb{Z})$. Upon recalling that $\|\cdot\|_{L^2(\mathcal{H}_\mathcal{Y})}$ satisfies the parallelogram identity and that G is convex, a quick computation shows

$$\|\mathbb{Z}_n - \mathbb{Z}_m\|^2_{L^2(\mathcal{H}_\mathcal{Y})} \;\leq\; 2\Phi(\mathbb{Z}_n) + 2\Phi(\mathbb{Z}_m) - 4\Phi(\mathbb{Y}^\Lambda_X) \quad\longrightarrow\quad 0 \quad (\text{for}\;\; n, m \to \infty), \tag{56}$$

which implies that $(\mathbb{Z}_k)_{k\in\mathbb{N}}$ is Cauchy. Hence, and since G is complete, there is $\mathbb{Z}_\star \in G$ such that $\mathbb{Z}_\star = \lim_{k\to\infty} \mathbb{Z}_k$ in $\|\cdot\|_{L^2(\mathcal{H}_\mathcal{Y})}$, whence we have $\Phi(\mathbb{Z}_\star) = \lim_{k\to\infty} \Phi(\mathbb{Z}_k) = \alpha$ and thus

$$\mathbb{Z}_\star \in \operatorname*{arg\,min}_{\mathbb{Z}\in G} \Phi(\mathbb{Z}), \quad \text{which, by (54), implies} \quad \mathbb{Z}_\star = \mathbb{Y}^\Lambda_X$$

(as noted in the lead-up to (53), the above arg min contains exactly one element only). This proves (50). Regarding the final claim on almost sure convergence, note $\|\mathbb{Z}_k - \mathbb{Z}_\star\|^2_{L^1(\mathcal{H}_\mathcal{Y})} \leq \|\mathbb{Z}_k - \mathbb{Z}_\star\|^2_{L^2(\mathcal{H}_\mathcal{Y})} \leq 2(\Phi(\mathbb{Z}_k) - \alpha) =: 2\beta_k$ by (56) [and since $\|\cdot\|_{L^1} \leq \|\cdot\|_{L^2}$], whence if $\sum_{k=0}^\infty \sqrt{\beta_k} < \infty$ then, by monotone convergence, $\int_\Omega \sum_{k=0}^\infty \|\mathbb{Z}_k - \mathbb{Z}_\star\| \,\mathrm{d}\mathbb{P} = \sum_{k=0}^\infty \|\mathbb{Z}_k - \mathbb{Z}_\star\|_{L^1(\mathcal{H}_\mathcal{Y})} < \infty$, thus $\sum_{k=0}^\infty \|\mathbb{Z}_k - \mathbb{Z}_\star\| < \infty$ $\mathbb{P}$-a.s. and hence $\lim_{k\to\infty} \|\mathbb{Z}_k - \mathbb{Z}_\star\| = 0$ a.s., as claimed. □

For the next corollary to [the proof of] Theorem 4.5, let $(e_i)_{i=1}^m$ be the standard basis of $\mathbb{R}^m$ and

$$\mathfrak{L}^2_X(\mathbb{R}^m) \;:=\; \big\{(\ell_1, \ldots, \ell_m) \in \big(\mathbb{R}[d]\big)^{\times m}\big\} \quad \text{and} \quad \psi^{[m]}_\alpha := \sum_{i=1}^m \langle \alpha_i, \underline{\mathfrak{sig}}_\Xi(\cdot)\rangle \cdot e_i \quad \text{for}\;\; \alpha \in \mathfrak{L}^2_X(\mathbb{R}^m).$$

The following is the first of the announced representations (2) and (3) from Section 1.

Corollary 4.6. *For any $Z \in L^2(\Omega, \mathscr{F}, \mathbb{P}; \mathbb{R}^m) \equiv L^2(\mathbb{R}^m)$, we have that*

$$\mathbb{E}\big[Z \,\big|\, X\big] \;=\; \lim_{k\to\infty} \psi^{[m]}_{\alpha_k}(X) \quad \textit{in} \;\; L^2_X(\mathbb{R}^m) \tag{57}$$

for any minimizing sequence $(\alpha_k) \subset \mathfrak{L}^2_X(\mathbb{R}^m)$ of the (convex) semi-infinite linear least-squares problem

$$\inf_{\alpha\in\mathfrak{L}^2_X(\mathbb{R}^m)} \mathbb{E}\big[\big|Z - \psi^{[m]}_\alpha(X)\big|^2\big]. \tag{58}$$

The convergence (57) holds $\mathbb{P}$-a.s. if the minimizing sequence (α_k) is such that $\sum_{k=0}^\infty(\Phi_m(\psi^{[m]}_{\alpha_k}(X)) - \eta)^{1/2} < \infty$, where $\Phi_m(\mathcal{W}) \coloneqq \mathbb{E}[\|Z - \mathcal{W}\|^2]$ and $\eta \coloneqq \inf_{\alpha\in\mathfrak{L}^2_X(\mathbb{R}^m)} \Phi_m(\psi^{[m]}_\alpha(X))$.

Proof. The proof of Theorem 4.5 remains valid as stated up to display (54) if we replace $(\mathbb{Y}^\Lambda, G)$ with $\big(Z, L^2_X(\mathbb{R}^m)\big)$, where $L^2_X(\mathbb{R}^m) \equiv \{\mathcal{W} \in L^2(\mathbb{R}^m) \mid \mathcal{W}$ is $(\Sigma_X, \mathcal{B}(\mathbb{R}^m))$-measurable$\}$. In particular

$$\mathbb{E}\big[Z \,\big|\, X\big] \;=\; \operatorname*{arg\,min}_{\mathcal{W}\in L^2_X(\mathbb{R}^m)} \;\mathbb{E}\big|Z - \mathcal{W}\big|^2, \tag{59}$$

and since the proof of Proposition 4.3 (cf. (48)) shows that $\mathfrak{H} \coloneqq \{\psi^{[m]}_\alpha(X) \mid \alpha \in \mathfrak{L}^2_X(\mathbb{R}^m)\}$ is an $\|\cdot\|_{L^2_X(\mathbb{R}^m)}$-dense subset of $L^2_X(\mathbb{R}^m)$, the corollary follows from (59) in the same way as Theorem 4.5 follows from (54). □

Remark 4.7. In many applications, the objectives in (51) and (58) can be approximated with a standard Monte Carlo average, that is with a empirical mean squared distances of the form

$$\widehat{\Phi}_N(\alpha) \;\coloneqq\; \frac{1}{N}\sum_{j=1}^N \big(\underline{\mathfrak{sig}}_\Lambda(y_j) - \tilde{\psi}_\alpha(x_j)\big)^2, \qquad \tilde{\psi} \in \{\psi, \psi^{[m]}\},$$

based on realisations (x_j, y_j) of (X, Y) that exhibit sufficiently weak internal statistical dependence. Associated questions of statistical consistency, along with how to handle necessary truncations of the [countably-infinite vectors which are the] appearing signatures, are to be expanded on in the full paper, cf. Section 6. ◆

5 Mathematical and Statistical Applications

We have introduced a way to variationally characterise the conditional expectation (50) of the signature cooredinates of a process (37) given the information (42) of another process. In this section, we apply this computationally feasible representation of conditional expectation to efficiently compute probabilistic quantities that are of interest in a variety of statistical and machine learning applications.

5.1 Computing Conditional Distributions

We start by showing how the conditional expectation $\mathbb{E}[\mathbb{Y}^\Lambda \,|\, X]$ and its estimates of (50) can be used to learn the conditional distribution (36). This states and shows the announced identity (3) for the conditional distributions of two stochastic processes.

Proposition 5.1. *For stochastic processes X and Y as in (34) and any Borel-set $A \subseteq \mathcal{Y}$, we have*

$$\mathbb{P}(Y \in A \mid X) \;=\; \lim_{l\to\infty} \big\langle \mathbb{E}\big[\mathbb{Y}^\Lambda \,\big|\, X\big],\, \ell^{(l)}_A \big\rangle \quad \textit{in} \;\; L^2_X(\mathbb{R}) \tag{60}$$

for any minimizing sequence $\big(\ell_A^{(l)} \,|\, l \in \mathbb{N}\big)$ *of the convex optimization problem*

$$\inf_{\ell \in \mathbb{R}[\tilde{d}]} \int_{\mathcal{Y}} \Big(\mathbb{1}_A(y) - \langle \ell, \underline{\mathfrak{sig}}_\Lambda(y)\rangle\Big)^2 \,\mathbb{P}_Y(\mathrm{d}y). \tag{61}$$

In particular, for $(\ell_A^{(l)})$ *as above and any sequence* $(\psi_{\alpha_k}(X))$ *as in* (50),

$$\mathbb{P}(Y \in A \mid X) \;=\; \lim_{l\to\infty}\lim_{k\to\infty} \big\langle \psi_{\alpha_k}(X),\, \ell_A^{(l)}\big\rangle \quad \textit{in} \;\; L^2_X(\mathbb{R}). \tag{62}$$

Proof. Fix any $A \in \mathcal{B}(\mathcal{Y})$. Due to (47) [replacing (X, Ξ) by (Y, Λ) in (46)] there is a sequence $(\ell_l) \subset \mathbb{R}[\tilde{d}]$ such that $\int_{\mathcal{Y}}(\mathbb{1}_A(y) - \langle \ell_l, \underline{\mathfrak{sig}}_\Lambda(y)\rangle)^2\,\mathbb{P}_Y(\mathrm{d}y) \;=\; \|\mathbb{1}_A(Y) - \langle \ell_l, \underline{\mathfrak{sig}}_\Lambda(Y)\rangle\|^2_{L^2(\Sigma_Y)} \to 0$ as $l \to \infty$, so the infimum (61) is zero. Hence for any fixed minimizing sequence $(\ell_A^{(l)}) \subset \mathbb{R}[\tilde{d}]$ of (61) we have

$$\begin{aligned}\big\| \mathbb{P}(Y \in A \mid X) - \mathbb{E}[\langle \ell_A^{(l)}, \mathbb{Y}^\Lambda\rangle \mid X]\big\|_{L^2(\Sigma_X)} \;&=\; \big\| \mathbb{E}[\mathbb{1}_A(Y) - \langle \ell_A^{(l)}, \mathbb{Y}^\Lambda\rangle \mid X]\big\|_{L^2(\Sigma_X)} \\ &\le\; \big\| \mathbb{1}_A(Y) - \langle \ell_A^{(l)}, \mathbb{Y}^\Lambda\rangle\big\|_{L^2(\Sigma_Y)} \longrightarrow 0 \quad \text{as } l \to \infty,\end{aligned} \tag{63}$$

where the last line is due to the Jensen's inequality (and the tower property of conditional expectations).

Now inherited from the analogous 'commuting' property of Bochner integrals, we have that

$$\mathbb{E}[\langle \ell_A^{(l)}, \mathbb{Y}^\Lambda\rangle \mid X] \;=\; \big\langle \ell_A^{(l)},\, \mathbb{E}[\mathbb{Y}^\Lambda \mid X]\big\rangle \qquad (\forall\, l \in \mathbb{N}) \tag{64}$$

with probability one, see for instance [36, Theorem II.2.3]. Combining (63) and (64) proves (60).

The convergence (62) is clear from (60) and (50). (Indeed: Since $\eta_k(\ell) \coloneqq \langle \psi_{\alpha_k}(X), \ell\rangle \in L^2_X(\mathbb{R})$ is merely a (finite) linear combination of [$L^2_X(\mathbb{R})$-valued] $\psi_{\alpha_k}(X)$-coefficients for any given $\ell \in \mathbb{R}[\tilde{d}]$, the $L^2_X(\mathbb{R})$-convergence $\langle \mathbb{E}[\mathbb{Y}^\Lambda \mid X], \ell\rangle = \lim_{k\to\infty} \eta_k(\ell)$ is readily implied by (50) [cf. (38)].) □

Approximations similar to (60) and (62) hold $\mathbb{P}$-a.s. if the law of Y is compactly supported. To prepare for this result, note that for $A \subset \mathcal{Y}$ open[9] the indicator $\mathbb{1}_A$ admits a monotone pointwise approximation by a sequence of nonnegative uniformly bounded functions in $C_b(\mathcal{Y})$; more specifically:

$$\mathbb{1}_A \uparrow h_A^{(\nu)} \quad (\nu \to \infty) \quad \text{pointwise,} \quad \text{e.g.} \quad h_A^{(\nu)} : y \mapsto \frac{\mathrm{d}(y, \mathcal{Y}\backslash A)}{\mathrm{d}(y, \mathcal{Y}\backslash A) + \mathrm{d}(y, F_\nu(A))} \in C_b(\mathcal{Y}), \tag{65}$$

where $\mathrm{d}(y, C) \coloneqq \inf_{z\in C}\|y - z\|_{1\text{-var}}$ (for $C \subset \mathcal{Y}$) and $F_\nu(A) \coloneqq \{y \in \mathcal{Y} \mid \mathrm{d}(y, \mathcal{Y}\backslash A) \ge \nu^{-1}\}$. From Proposition 3.15, eq. (33), we know that the continuous approximants $h_A^{(\nu)}$ can each be written as

$$h_A^{(\nu)} \;=\; \lim_{\mu\to\infty} \big\langle \ell_{A_\nu}^{(\mu)}, \underline{\mathfrak{sig}}_\Lambda(\,\cdot\,)\big\rangle \;\; \text{wrt.} \;\; \tau^{\mathcal{Y}}_{\text{str}}, \quad \text{for some} \;\; \big(\ell_{A_\nu}^{(\mu)}\big)_{\mu\in\mathbb{N}} \subset \mathbb{R}[\tilde{d}]. \tag{66}$$

Corollary 5.2. *Let* X *and* Y *be as in* (34) *but with* $\mathrm{supp}(\mathbb{P}_Y)$ *compact. Then for* $A \subseteq \mathcal{Y}$ *open or closed,*

$$\mathbb{P}(Y \in A \mid X) \;=\; \lim_{\nu\to\infty}\lim_{\mu\to\infty} \big\langle \mathbb{E}\big[\mathbb{Y}^\Lambda \,\big|\, X\big],\, \ell_{A_\nu}^{(\mu)}\big\rangle \quad \mathbb{P}\textit{-a.s.} \tag{67}$$

for any $\big(\ell_{A_\nu}^{(\mu)}\big)_{\mu,\nu}$ *as in* (66)*. Moreover, for any sequence* $(\psi_{\alpha_k}(X))$ *as in* (50) *we have*

$$\mathbb{P}(Y \in A \mid X) \;=\; \lim_{\nu\to\infty}\lim_{\mu\to\infty}\lim_{k\to\infty} \big\langle \psi_{\alpha_k}(X),\, \ell_{A_\nu}^{(\mu)}\big\rangle \quad \textit{in} \;\; L^2_X(\mathbb{R}) \tag{68}$$

and the convergence in (68) *holds* $\mathbb{P}$*-a.s. if* $(\psi_{\alpha_k}(X))$ *is chosen such that* (50) *converges almost surely.*

[9] Since $\mathbb{1}_A = 1 - \mathbb{1}_{A^c}$, an analogous approximation of $\mathbb{1}_A$ can be found if the set A is closed.

Proof. Fix any open $A \subseteq \mathcal{Y}$. (Since $\mathbb{P}(Y \in A \mid X) = 1 - \mathbb{P}(Y \in A^c \mid X)$, the case for closed A is contained herein.) From the preliminary observations (65) and (66) and the definition of cond. probab., cf. (36),

$$\mathbb{P}(Y \in A \,|\, X) \;=\; \mathbb{E}\big[\mathbb{1}_A(Y) \,\big|\, X\big] \;=\; \lim_{\nu\to\infty} \mathbb{E}\big[h_A^{(\nu)}(Y) \,\big|\, X\big] \quad \mathbb{P}\text{-a.s.} \tag{69}$$

via the conditional monotone convergence theorem. Now since $\mathfrak{D}_Y \coloneqq \operatorname{supp}(\mathbb{P}_Y)$ is assumed compact,

$$\forall\,\nu \in \mathbb{N} \,:\, \exists\,\big(\ell_\nu^{(\mu)}\big)_\mu \quad \text{s.t.} \quad \lim_{\mu\to\infty} \big\|h_A^{(\nu)} - \langle \ell_\nu^{(\mu)}, \underline{\mathfrak{sig}}_\Lambda\rangle\big\|_{\infty;\mathfrak{D}_Y} \;=\; 0 \tag{70}$$

by (66). Now for $\nu \in \mathbb{N}$ fixed, take any $(\ell_\nu^{(\mu)})_\mu$ as in (70) and denote $\varsigma_\mu(y) \coloneqq h_A^{(\nu)}(y) - \langle \ell_\nu^{(\mu)}, \underline{\mathfrak{sig}}_\Lambda(y)\rangle$. Since $\mathbb{P}(Y \in \mathfrak{D}_Y) = 1$, or even $Y(\Omega) \subseteq \mathfrak{D}_Y$ wlog (Remark A.1 (iii)), we have

$$C \coloneqq \sup_{\mu\in\mathbb{N}} \sup_{\omega\in\Omega} \big|\varsigma_\mu(Y(\omega))\big| \;<\; \infty.$$

Indeed, given $\epsilon > 0$ there is $\mu_0 \in \mathbb{N}$ with $\sup_{\mu\geq\mu_0,\omega} |\varsigma_\mu(Y(\omega))| \leq \sup_{\mu\geq\mu_0} \|\varsigma_\mu\|_{\infty;\mathfrak{D}_Y} \leq \epsilon$ by (70), and hence $C \;\leq\; \max_{\mu\leq\mu_0} \|\varsigma_\mu\|_{\infty;\mathfrak{D}_Y} + \epsilon \;\leq\; 1 + \|\Lambda\|_{\infty;\mathcal{H}} \cdot \max_{\mu\leq\mu_0} \|\ell_\nu^{(\mu)}\| + \epsilon \;<\; \infty$.

Now clearly $\sup_\mu |\varsigma_\mu(Y)| \leq C$ a.s. as well as $|\varsigma_\mu(Y)| \leq \|\varsigma_\mu\|_{\infty;\mathfrak{D}_Y} \to 0$ $(\mu\to\infty)$ a.s., and hence

$$\Big|\mathbb{E}\big[h_A^{(\nu)}(Y) \,\big|\, X\big] - \mathbb{E}\big[\langle \ell_\nu^{(\mu)}, \underline{\mathfrak{sig}}_\Lambda(Y)\rangle \,\big|\, X\big]\Big| \;=\; \big|\mathbb{E}[\varsigma_\mu(Y) \,|\, X]\big| \;\overset{\mu\to\infty}{\longrightarrow}\; 0 \quad \mathbb{P}\text{-a.s.} \tag{71}$$

via the conditional dominated convergence theorem. Combining (69) and (71) and (64) yields (67).

The convergence (68) follows as in (62), while the corollary's last assertion is an immediate consequence of (67) and (50) and the continuous mapping theorem. □

Remark 5.3. Suitable $\underline{\mathfrak{sig}}_\Lambda$-discretizations $(\ell_A^{(\mu)})$ of an indicator $\mathbb{1}_A$ as in (66) can be found approximately by solving (for $\boldsymbol{u}$, given $\boldsymbol{S}$, $\boldsymbol{b}_A$) the (semi-infinite) linear least squares problem [41]

$$\|\boldsymbol{S}\boldsymbol{u} - \boldsymbol{b}_A\|_2 \overset{!}{=} \min \quad \text{for} \quad \boldsymbol{S} \coloneqq \begin{pmatrix} \underline{\mathfrak{sig}}_\Lambda(y_1) \\ \vdots \\ \underline{\mathfrak{sig}}_\Lambda(y_n) \end{pmatrix} \quad \text{and} \quad \boldsymbol{b}_A \coloneqq \begin{pmatrix} \mathbb{1}_A(y_1) \\ \vdots \\ \mathbb{1}_A(y_n) \end{pmatrix},$$

were $(y_j)_{j=1}^n$, $n \in \mathbb{N}$, are some observations of Y. (Of course, in practice the matrix $\boldsymbol{S}$ will be replaced by a truncation $\boldsymbol{S}_m \coloneqq \big(\pi_{[m]}(\underline{\mathfrak{sig}}_\Lambda(y_1))| \cdots |\pi_{[m]}(\underline{\mathfrak{sig}}_\Lambda(y_n))\big)^\intercal$ at some cutoff-level $m \in \mathbb{N}$.) ◆

5.2 Stochastic Process Prediction

Another application of common interest is to predict the future evolution of a stochastic process given information on its past. As a potential scenario for this, we may specialise our general setting (34) to

$$Y \coloneqq \big(X_{t\wedge t_f}\big)_{t\in[0,1]} \quad \text{and} \quad X \coloneqq \big(X_{t\wedge t_p}\big)_{t\in[0,1]} \quad \text{for some} \quad 0 \;<\; t_p \;<\; t_f \;<\; 1.$$

The X-generated Borel σ-algebra $\Sigma_X = \sigma(X)$ then reads (cf. [20, Problem 2.4.4.2, p. 60])

$$\Sigma_X \;=\; \sigma(X_s \mid 0 \leq s \leq t_p),$$

and the ultimate goal is to find the '(L^2-)best-prediction' of the 'full-scale' process $X|_{[0,t_f]}$ (evolving up to some time-horizon t_f in the future) given the knowledge $X|_{[0,t_p]}$ of said process over a historical time-window $[0,t_p]$. This goal is uniquely achieved by the conditional expectation $\mathbb{E}[Y \,|\, X]$.

While not providing a construction for the path-valued object $\mathbb{E}[Y\,|\,X]$ directly, Theorem 4.5 gives us an algorithm to compute the L^2-projection $\mathbb{E}[\mathbb{Y}^\Lambda\,|\,X]$ of the (normalised) *Hilbert coordinates* (40) of Y given X. As stated in Proposition 5.1 and Corollary 5.2, this construction allows us to compute the conditional distribution of the full process Y given our knowledge of its history X. If instead of its full trajectories we are only interested in the values of Y at finitely many time-points, say at $(t_j)_{j=1}^N \subset (t_p\,, t_f]$, then these can be 'best-predicted' with Corollary 4.6 applied to $Z \coloneqq (Y_{t_1}, \cdots, Y_{t_N})$.

6 Outlook

As mentioned in the introduction, this document is only a working paper to present some mathematical core ideas on how to condition on and between stochastic processes using their signatures. The full version of this paper, which is currently still work in progress, is intended to contain mathematical extensions of the present core results (that is, of Theorem 4.5, Corollary 4.6, and Proposition 5.1), ideally including complementary statements on statistical consistency and convergence rates, as well as numerical examples and the incorporation of a detailed application to illustrate the practical relevance of the presented conditioning theory more fully.

Appendix A Some Remarks on Stochastic Processes

Remark A.1. Writing $\mathbb{P}_X \coloneqq \mathbb{P}\circ X^{-1}$ and $X(\omega) \equiv (X_t(\omega))_{t\in[0,1]}$ for each $\omega \in \Omega$, let us note the following facts on the conditioning process X from (34).

(i) The rv X is $\big(\mathscr{F}, \mathcal{B}(\|\cdot\|_{\text{1-var}})\big)$-measurable by def. Now $\mathcal{B}(\|\cdot\|_{\text{1-var}}) = \mathcal{B}(\|\cdot\|_\infty)$ by Lemma 3.1, where $\mathcal{B}(\|\cdot\|_\infty) = \sigma(\pi_t\,|\,t\in[0,1]) =: \mathcal{B}(\mathcal{C}^1_{d_\mathcal{X}}) = \mathcal{B}(\mathcal{C}_{d_\mathcal{X}}) \cap \mathcal{C}^1_{d_\mathcal{X}}$ is the Borel σ-algebra on $(\mathcal{C}^1_d, \|\cdot\|_\infty)$. (Here, $\pi_t : (x_t)_{t\in[0,1]} \mapsto x_t$ is the t-projection from $\mathcal{C}_{d_\mathcal{X}}$ onto $\mathbb{R}^{d_\mathcal{X}}$.) Consequently,

$$\mathbb{P}_X \in \mathcal{M}_1\big(\mathcal{C}^1_{d_\mathcal{X}}\,, \|\cdot\|_\infty\big) \quad \text{and} \quad X_t : \Omega \ni \omega \mapsto X_t(\omega) \in \mathbb{R}^{d_\mathcal{X}} \ \text{ is } \ \big(\mathscr{F}, \mathcal{B}(\mathbb{R}^{d_\mathcal{X}})\big)\text{-measurable}$$

for each $t \in [0,1]$. Hence, we can equivalently define the stochastic process $X : \Omega \to \mathcal{X}$ as an $[0,1]$-indexed family $(X_t)_{t\in[0,1]}$ of (Borel) random vectors $X_t : \Omega \to \mathbb{R}^{d_\mathcal{X}}$ such that $t \mapsto X_t(\omega)$ is continuous for each $\omega \in \Omega$, see e.g. [35, Section II.27].

(ii) In stricter terminology, a stochastic process $X \equiv (X_t(\omega)) : [0,1]\times\Omega \to \mathbb{R}^d$ defined as a $(\mathscr{F}, \mathcal{B}(\mathcal{C}^1_{d_\mathcal{X}})$-measurable map $X : \Omega \to \mathcal{C}_{d_\mathcal{X}}$, as we did above, is called *jointly measurable*. If $(\Omega, \mathscr{F}, \mathbb{P})$ is filtered then it can carry stronger measurability notions (such as progressive measurability or predictability, see e.g. [44, Proposition 2.23]), but for our purposes the weak notion of joint measurability will suffice.

(iii) It will be no loss of generality for us to assume (if convenient) that in fact

$$X : \Omega \to \mathfrak{D}_X\,, \quad \text{where} \quad \mathfrak{D}_X \coloneqq \operatorname{supp}(\mathbb{P}_X)$$

is the support of X. Indeed: By its definition, the support $\mathfrak{D}_X$ is the smallest closed subset $C \subseteq \mathcal{X}$ for which $\mathbb{P}_X(C) = 1$, see e.g. [19, Lemma 1.19]. Hence $\tilde{\Omega} \coloneqq X^{-1}(\mathfrak{D}_X) \in \mathscr{F}$ is a $\mathbb{P}$-full set, which implies that X and its $\mathfrak{D}_X$-valued twin $\tilde{X} \coloneqq \mathbb{1}_{\tilde{\Omega}} \cdot X + \mathbb{1}_{\Omega\setminus\tilde{\Omega}} \cdot x_0 : \Omega \to \mathfrak{D}_X$ (any $x_0 \in \mathfrak{D}_X$ fixed) are indistinguishable.

(iv) We will further assume that the X-induced sub-σ-algebra $\Sigma_X \coloneqq \sigma(X) \subseteq \mathscr{F}$ is $\mathbb{P}$-complete. Recall that this assumption entails no loss of generality: If $(\Omega, \Sigma_X, \mathbb{P})$ is not complete, we can immediately and 'minimally' complete it as follows. Writing $\mathcal{N}_\mathbb{P} \coloneqq \{N \subseteq \Omega \mid \exists\, A \in$

$\mathscr{F} : N \subseteq A$ and $\mathbb{P}(A) = 0\}$ for the system of all subsets of $\mathbb{P}$-nullsets [in $\mathscr{F}$], define $\Sigma_X^{\mathbb{P}} \coloneqq \{A \cup N \mid A \in \Sigma_X,\ N \in \mathcal{N}_{\mathbb{P}}\}$ and $\mathscr{F}^{\mathbb{P}} \coloneqq \{A \cup N \mid A \in \mathscr{F},\ N \in \mathcal{N}_{\mathbb{P}}\}$ and $\bar{\mathbb{P}} : \mathscr{F}^{\mathbb{P}} \to [0,1]$ by $\bar{\mathbb{P}}(A \cup N) \coloneqq \mathbb{P}(A)$ for all $A, \in \mathscr{F}, N \in \mathcal{N}_{\mathbb{P}}$. Then $\Sigma_X^{\mathbb{P}} \subseteq \mathscr{F}^{\mathbb{P}}$, both $(\Omega, \Sigma_X^{\mathbb{P}}, \bar{\mathbb{P}})$ and $(\Omega, \mathscr{F}^{\mathbb{P}}, \bar{\mathbb{P}})$ are complete probability spaces, and each complete extension μ of $\mathbb{P}$ is an extension of $\bar{\mathbb{P}}$, e.g. [13, Satz 6.3]. All our objects of interest stay the same when passing to this completion, that is (trivially) $\mathbb{E}[Y \mid \Sigma_X] = \mathbb{E}[Y \mid \Sigma_X^{\mathbb{P}}]$ $\bar{\mathbb{P}}$-a.s. and $L^p(\Omega, \Sigma_X, \mathbb{P}; \mathcal{H}) \cong L^p(\Omega, \Sigma_X^{\mathbb{P}}, \bar{\mathbb{P}}; \mathcal{H})$ (canonically). ◆

Appendix B Additional Proofs

B.1 Proof of Lemma 3.1

Proof. The first assertion holds by [14, Propositions 1.31 & 1.32]. In fact, [14, Prop. 1.31] asserts that for $\mathfrak{X} \coloneqq \mathbb{R}^d \times L^1([0,1];\mathbb{R}^d)$ and $\mathfrak{Y} \coloneqq \mathcal{C}_d$, the map $f : \mathfrak{X} \to \mathfrak{Y}$ given by $f(c,v) \coloneqq c + \int_0^{\cdot} v_s \, \mathrm{d}s$ is a Banach space isomorphism (which also proves the norm identity $\|x\|_{1\text{-var}} = |x_0| + \|\dot{x}\|_{L^1}$ on $\mathcal{C}_d^1$). From this, [21, Theorem 15.1] implies that the image $f(\mathfrak{X}) = \mathcal{C}^1$ is a Borel subset of $(\mathcal{C}^1, \|\cdot\|_\infty)$, i.e. that $\mathcal{C}^1 \in \mathcal{B}(\mathcal{C}_d)$. That $(\mathcal{C}^1, \|\cdot\|_{\text{-var}})$ is separable and Banach is stated as [14, Corollary 1.35].

For the lemma's second assertion, note first that since $\|\cdot\|_{1\text{-var}} \geq \|\cdot\|_\infty$ (which is easy to see), we find that the 1-variation topology on $\mathcal{C}^1$ is finer than the uniform topology on $\mathcal{C}^1$, which of course implies that $\mathcal{B}_{1\text{-var}} \coloneqq \sigma(\mathcal{C}^1, \|\cdot\|_{1\text{-var}}) \supseteq \sigma(\mathcal{C}^1, \|\cdot\|_\infty) =: \mathcal{B}_\infty$. Since the separability of $(\mathcal{C}^1, \|\cdot\|_{1\text{-var}})$ guarantees that the σ-algebra $\mathcal{B}_{1\text{-var}}$ is generated by the closed $\|\cdot\|_{1\text{-var}}$-balls, the converse inclusion $\mathcal{B}_{1\text{-var}} \subseteq \mathcal{B}_\infty$ follows if we can show that

$$B_r^1(x) \coloneqq \{y \in \mathcal{C}^1 \mid \|y - x\|_{1\text{-var}} \leq r\} \ \in \ \mathcal{B}_\infty \quad \text{for every} \quad x \in \mathcal{C}^1 \text{ and any } r \geq 0. \tag{72}$$

To see that this holds, fix any $x \in \mathcal{C}^1$ and $r \geq 0$ and recall that, by definition of the 1-variation norm,

$$\|z\|_{1\text{-var}} = \sup_{\mathcal{I} \in \mathfrak{I}} V_{\mathcal{I}}(z) \quad \text{with} \quad V_{(t_\nu)}(z) \coloneqq |z_0| + \sum_\nu \left| z_{t_{\nu+1}} - z_{t_\nu} \right|$$

and where $\mathfrak{I} \coloneqq \{\mathcal{I} = (t_\nu) \mid \mathcal{I} \text{ is a (finite) dissection of } [0,1]\}$. Given any $\mathcal{I} \in \mathfrak{I}$ it is clear that the function $Q_{\mathcal{I}} : \mathcal{C}^1 \ni y \mapsto V_{\mathcal{I}}(y - x)$ is continuous wrt. $\|\cdot\|_\infty$, whence the level set $C_{\mathcal{I}} \coloneqq \{y \in \mathcal{C}^1 \mid Q_{\mathcal{I}}(y) \leq r\}$ is $\|\cdot\|_\infty$-closed. Combined with this, the immediate identity

$$B_r^1(x) = \bigcap\nolimits_{\mathcal{I} \in \mathfrak{I}} C_{\mathcal{I}} \quad \text{implies that} \quad B_r^1(x) \text{ is closed wrt. } \|\cdot\|_\infty\,,$$

which shows that (72) holds as desired. □

B.2 Proof of Lemma 3.2

Proof. Note that since both $(\mathcal{X}, \|\cdot\|_{1\text{-var}})$ and $(\mathcal{Y}, \|\cdot\|_{1\text{-var}})$ are Polish by Lemma 3.1, so is the product space $(\mathcal{X} \times \mathcal{Y}, \|\cdot\|_\alpha)$ with $\|(x,y)\|_\alpha \coloneqq \max\{\|x\|_{1\text{-var}}, \|y\|_{1\text{-var}}\}$, as the norm $\|\cdot\|_\alpha$ induces the product topology on $(\mathcal{X}, \|\cdot\|_{1\text{-var}}) \times (\mathcal{Y}, \|\cdot\|_{1\text{-var}})$. This implies that $\mathcal{Z}$ itself is Polish, since the norms $\|\cdot\|_\alpha$ and $\|\cdot\|_{1\text{-var}}$ are equivalent on $\mathcal{X} \times \mathcal{Y}$. The latter equivalence of norms also gives that

$$\mathcal{B}(\mathcal{Z}, \|\cdot\|_{1\text{-var}}) = \mathcal{B}(\mathcal{Z}, \|\cdot\|_\alpha), \tag{73}$$

and since further $\mathcal{B}(\mathcal{Z}, \|\cdot\|_\alpha) = \mathcal{B}(\mathcal{X}, \|\cdot\|_\infty) \otimes \mathcal{B}(\mathcal{Y}, \|\cdot\|_\infty) = \mathcal{B}(\mathcal{Z}, \|\cdot\|_\beta)$ for the norm $\|z\|_\beta \coloneqq \max\{\|\pi_{\mathcal{X}}(z)\|_\infty, \|\pi_{\mathcal{Y}}(z)\|_\infty\}$ by Lemma 3.1 (recalling that: (a) the Borel σ-algebra of the product of two (second countable) topological spaces equals the product of their Borel σ-algebras, and (b) the norm $\|\cdot\|_\beta$ induces the product topology on $(\mathcal{X}, \|\cdot\|_\infty) \times (\mathcal{Y}, \|\cdot\|_\infty)$) and with the norms $\|\cdot\|_\infty$ and $\|\cdot\|_\beta$ being equivalent on $\mathcal{X} \times \mathcal{Y}$, we find that $\mathcal{B}(\mathcal{Z}, \|\cdot\|_{1\text{-var}}) = \mathcal{B}(\mathcal{Z}, \|\cdot\|_\infty)$ as desired.

The claimed characterisation of measurability holds by (73) (upon recalling that $\mathcal{B}(\mathcal{Z}, \|\cdot\|_\alpha) = \mathcal{B}(\mathcal{X}, \|\cdot\|_{\text{1-var}}) \otimes \mathcal{B}(\mathcal{Y}, \|\cdot\|_{\text{1-var}})$) and the fact that a product-space-valued function (here: Z) is product-measurable iff all of its factor components (here: X and Y) are measurable. □

B.3 Proof of Lemma 3.6

Proof. Each of the spaces $(V_m, \langle\cdot,\cdot\rangle_m)$ from (19) is Hilbert and separable (with ONB $[\boldsymbol{d}]^*_m := \{w \in [\boldsymbol{d}]^* \mid |w| = m\}$). Thus the space $\mathcal{H}$ is Hilbert and separable — with ONB $[\boldsymbol{d}]^*$ — as the Hilbert direct sum of the family $\{(V_m, \langle\cdot,\cdot\rangle_m) \mid m \in \mathbb{N}_0\}$, see e.g. [12, Proposition I.6.2]. The inclusion $\mathfrak{sig}(\mathcal{C}^1_{\boldsymbol{d}}) \subset \mathcal{H}$ follows from the factorial decay of the signature coefficients, cf. [9, Corollary 5.5].

The last assertion follows from the usual p-variation continuity of $\mathfrak{sig}$, see e.g. [9, Corollary 5.5], and the fact that the locally convex topology from [9, Section 2] (defined by the (fundamental) family of semi-norms $\Psi := (|\!|\!|\cdot|\!|\!|_\lambda \mid \lambda > 0)$ on V, where $|\!|\!|\boldsymbol{t}|\!|\!|_\lambda := \sum_{m\geq 0} \|\pi_m(\boldsymbol{t})\|_m \cdot \lambda^m$; denote the associated locally m-convex topology by τ_{lc}) is finer than the [canonical, i.e. $\|\cdot\|$-induced] topology on $\mathcal{H}$ (denote this topology by $\tau_{\mathcal{H}}$). To prove the asserted inclusion of topologies: Since τ_{lc} is metrizable, see e.g. [9, Corollary 2.4], the topological space (V, τ_{lc}) is sequential, whence $\tau_{\mathcal{H}} \subseteq \tau_{\text{lc}}$ iff every τ_{lc}-convergent sequence in V is $\tau_{\mathcal{H}}$ convergent. This clearly holds, however, since for every null-sequence (v_k) in (V, τ_{lc}) there is $k_0 \in \mathbb{N}$ with $\sup_{k \geq k_0} |\!|\!|v_k|\!|\!|_\lambda < 1$ (for some $\lambda > 1$), whence for $k \geq k_0$ we find that $\|v_k\|^2 = \sum_{m\geq 0} \|\pi_m(v_k)\|^2_m \leq |\!|\!|v_k|\!|\!|_\lambda$ goes to zero as $k \to \infty$. □

B.4 Proof of Proposition 3.15

Proof. Let us first note that, clearly,

$$\underline{\mathfrak{sig}}_\Lambda = \Lambda \circ \underline{\mathfrak{sig}}. \tag{74}$$

Indeed, $\pi_m(\underline{\mathfrak{sig}}_\Lambda(x)) = \underline{\lambda}^m_x \sum_{|w|=m} \xi_w(\bar{x}) = \lambda^m_{\underline{\mathfrak{sig}}(x)}\big[\big(\sum_{|w|=m} \xi_w\big) \circ \bar{\iota}\big](x) = \lambda^m_{\underline{\mathfrak{sig}}(x)}\big[\pi_m \circ \underline{\mathfrak{sig}}\big](x) = \pi_m\big(\lambda^m_{\underline{\mathfrak{sig}}(x)} \cdot \pi_m(\underline{\mathfrak{sig}}(x))\big) = \pi_m\big(\delta_{\lambda_{\underline{\mathfrak{sig}}(x)}}(\underline{\mathfrak{sig}}(x))\big) = \pi_m\big((\Lambda \circ \underline{\mathfrak{sig}})(x)\big)$ for each $(x, m) \in \mathcal{X} \times \mathbb{N}_0$.

To see that $\mathcal{A}_\Lambda \subset C(\mathcal{X})$, note [from (32)] that since each function $\varphi \in \mathcal{A}_\Lambda$ can be represented as

$$\varphi = \langle \ell_\varphi, \underline{\mathfrak{sig}}_\Lambda \rangle \quad \text{for some} \quad \ell_\varphi \in \mathbb{R}[d_0],$$

the desired $\|\cdot\|_{\text{1-var}}$-continuity of φ follows from (74) and the continuity assertion (27) of Lem. 3.13.

Since for $\mathbf{1} = 1 \cdot \emptyset \in \mathbb{R}[d_0]$ we have $\underline{\xi}^\lambda_\emptyset = \langle \mathbf{1}, \underline{\mathfrak{sig}}_\Lambda \rangle = \lambda^0_{\mathbf{1}} \cdot \langle \emptyset, \underline{\mathfrak{sig}} \rangle \equiv 1$ on $\mathcal{X}$, clearly $\mathcal{A}_\Lambda$ is non-vanishing. For $\mathcal{A}_\Lambda$ being point-separating, note that for any $x, y \in \mathcal{X}$ with $x \neq y$ we have $\underline{\mathfrak{sig}}(x) \neq \underline{\mathfrak{sig}}(y)$ by Lemma 18, and hence also $\underline{\mathfrak{sig}}_\Lambda(x) \neq \underline{\mathfrak{sig}}_\Lambda(y)$ by (74) and the injectivity of Λ. This implies that there is $w_0 \in [d_0]^*$ such that $\langle w_0, \underline{\mathfrak{sig}}_\Lambda(x) \rangle \neq \langle w_0, \underline{\mathfrak{sig}}_\Lambda(y) \rangle$, whence for $\varphi := \underline{\xi}^\lambda_{w_0} \in \mathcal{A}_\Lambda$ we find $\varphi(x) \neq \varphi(y)$.

To prove that $\mathcal{A}_\Lambda$ is an algebra, we need to show $\varphi \cdot \psi \in \mathcal{A}_\Lambda$ for any two $\varphi, \psi \in \mathcal{A}_\Lambda$. And indeed,

$$\begin{aligned}
\varphi \cdot \psi &= \sum_{w, \tilde{w} \in [d_0]^*} \langle \ell_\varphi, w \rangle \langle \ell_\psi, \tilde{w} \rangle \underline{\lambda}_{\cdot}^{|w|} \underline{\lambda}_{\cdot}^{|\tilde{w}|} \langle w, \underline{\mathfrak{sig}} \rangle \langle \tilde{w}, \underline{\mathfrak{sig}} \rangle & (75)\\
&= \sum_{w, \tilde{w} \in [d_0]^*} \underline{\lambda}_{\cdot}^{|w|+|\tilde{w}|} \langle \ell_\varphi, w \rangle \langle \ell_\psi, \tilde{w} \rangle \langle w \sqcup\!\sqcup \tilde{w}, \underline{\mathfrak{sig}} \rangle & (76)\\
&= \sum_{m \geq 0} \underline{\lambda}_{\cdot}^{m} \sum_{w, \tilde{w} \,:\, |w \sqcup\!\sqcup \tilde{w}| = m} \langle \ell_\varphi, w \rangle \langle \ell_\psi, \tilde{w} \rangle \langle w \sqcup\!\sqcup \tilde{w}, \underline{\mathfrak{sig}} \rangle & (77)\\
&= \sum_{m \geq 0} \underline{\lambda}_{\cdot}^{m} \langle \pi_m(\ell_{\varphi\psi}), \underline{\mathfrak{sig}} \rangle & (78)\\
&= \langle \ell_{\varphi\psi}, \Lambda \circ \underline{\mathfrak{sig}} \rangle, \quad \text{for} \quad \ell_{\varphi\psi} := \textstyle\sum_{w, \tilde{w} \in [d_0]^*} \langle \ell_\varphi, w \rangle \langle \ell_\psi, \tilde{w} \rangle \cdot w \sqcup\!\sqcup \tilde{w}. & (79)
\end{aligned}$$

Since both ℓ_φ, $\ell_\psi \in \mathbb{R}[d_0]$ (and thus $\langle \ell_\varphi, w\rangle = 0$ and $\langle \ell_\psi, w\rangle = 0$ for almost all $w \in [d_0]^*$), we get that also $\ell_{\varphi\psi} \in \mathbb{R}[d_0]$ and hence $\varphi \cdot \psi \in \mathcal{A}_\Lambda$ as desired. A few remarks on this are in order:

While (75) holds simply by linearity of (20), equation (76) involved the *character identity*

$$\langle \ell_1, \underline{\mathfrak{sig}}\rangle \cdot \langle \ell_2, \underline{\mathfrak{sig}}\rangle \;=\; \langle \ell_1 \sqcup\!\sqcup\, \ell_2, \underline{\mathfrak{sig}}\rangle, \quad \text{for any} \;\; \ell_1, \ell_2 \,\in\, \mathbb{R}[d_0]$$

(see e.g. [29, proof of Thm. 2.15]), where $\sqcup\!\sqcup : \mathbb{R}[d_0]^{\times 2} \to \mathbb{R}[d_0]$ is the so-called *shuffle product*, defined e.g. in [29, eq. (2.5) (p. 35)]. Denoting by $|\ell| := \max\{|w| \mid w \in [d_0]^* : w \in \ell\}$ the maximal length of any word contained (as a summand) in a given polynomial $\ell \in \mathbb{R}[d_0]$, it holds that $|\ell_1 \sqcup\!\sqcup\, \ell_2| = |\ell_1| + |\ell_2|$. This justifies equation (77), where we also used that $[d_0]^* \times [d_0]^* = \bigsqcup_{m\geq 0}\{(w, \tilde{w}) \in [d_0]^* \times [d_0]^* \mid |w \sqcup\!\sqcup\, \tilde{w}| = m\}$ defines a (disjoint) partition. For equation (78) we used that $\pi_m(\ell_{\varphi\psi}) = \sum_{|w\sqcup\!\sqcup\tilde{w}|=m} \langle \ell_\varphi, w\rangle\langle \ell_\psi, \tilde{w}\rangle \, w \sqcup\!\sqcup\, \tilde{w}$, and the concluding identity (79) follows from $\langle \ell_{\varphi\psi}, \Lambda \circ \underline{\mathfrak{sig}}\rangle = \sum_{m\geq 0} \underline{\lambda}_{\cdot}^{m} \langle \ell_{\varphi\psi}, \pi_m(\underline{\mathfrak{sig}})\rangle$ upon noting that $\langle \ell_{\varphi\psi}, \pi_m(\underline{\mathfrak{sig}})\rangle = \langle \pi_m(\ell_{\varphi\psi}), \pi_m(\underline{\mathfrak{sig}})\rangle = \langle \pi_m(\ell_{\varphi\psi}), \underline{\mathfrak{sig}}\rangle$ for each $m \geq 0$ (which is a trivial consequence of the definition (20) of $\langle\cdot,\cdot\rangle$).

We thus saw that $\mathcal{A}_\Lambda$ is a subalgebra of $C(\mathcal{X})$. Now if Λ is in fact an fN of the form (31), then

$$\|\varphi\|_\infty := \sup_{x\in\mathcal{X}} \left|\langle \ell_\varphi, \underline{\mathfrak{sig}}_\Lambda(x)\rangle\right| \;\leq\; \|\ell_\varphi\| \sup \|\Lambda(\underline{\mathfrak{sig}}(\mathcal{X}))\| \;\leq\; \|\ell_\varphi\| R \;<\; \infty$$

by Cauchy-Schwarz, which shows that in this case even $\mathcal{A}_\Lambda \subset C_b(\mathcal{X})$ as claimed.

The asserted denseness of $\mathcal{A}_\Lambda$ in $(C_b(\mathcal{X}), \tau^{\mathcal{X}}_{\text{str}})$ is then guaranteed by [15, Theorem 3.1], which generalises the theorem of Stone-Weierstrass to the $\tau^{\mathcal{X}}_{\text{str}}$-modulated non-compact setting.
If finally we are in the (unnormalised) special case $\Lambda = \mathrm{id}_{\mathcal{H}_\downarrow}$, that is if $\lambda_{\cdot} \equiv 1$, then the above arguments show that $\mathcal{A} := \mathcal{A}_{\mathrm{id}_{\mathcal{H}_\downarrow}}$ is a subalgebra of $C(\mathcal{X})$, while the bounds (24) yield that, for each $\varphi \in \mathcal{A}$,

$$\|\varphi\|_{\infty;\mathcal{X}} \;=\; \left\|\langle \ell_\varphi, \underline{\mathfrak{sig}}\rangle\right\|_{\infty;\mathcal{X}} \;\leq\; \|\ell_\varphi\| \sup\nolimits_{x\in\mathcal{X}} \|\underline{\mathfrak{sig}}(x)\| \;\leq\; \|\ell_\varphi\| \sup_{x\in\mathcal{X}} \sum_{m\geq 0} \frac{\|\bar{x}\|^m_{1\text{-var}}}{m!} \;\leq\; \|\ell_\varphi\| e^{\kappa_{\mathcal{X}}+1} \;<\; \infty$$

if $\kappa_{\mathcal{X}} := \sup_{x\in\mathcal{X}} \|x\|_{1\text{-var}}$ is assumed finite, and then $\mathcal{A} \subset C_b(\mathcal{X})$ as desired. □

References

[1] R. Abergel, C. Louchet, L. Moisan, and T. Zeng. *Total Variation Restoration of Images Corrupted by Poisson Noise with Iterated Conditional Expectations.* Springer, 2015.

[2] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent Measures of Risk. *Math. Finance*, **9**.3:203–228, 1999.

[3] A. Bain and D. Crisan. *Fundamentals of Stochastic Filtering.* Stochastic Modelling and Applied Probability **3**, Springer, 2009.

[4] C. Bayer, L. Pelizzari, and J. Schoenmakers. Primal and dual optimal stopping with signatures. *arXiv preprint arXiv:2312.03444*, 2023.

[5] A. Borovykh, S. Bohte, and C.W. Oosterlee. Conditional Time Series Forecasting with Convolutional Neural Networks. *arXiv preprint arXiv:1703.04691*, 2017.

[6] H.-P. Breuer, E.-M. Laine, J. Piilo, and B. Vacchini. Colloquium: Non-Markovian Dynamics in Open Quantum Systems. *Reviews of Modern Physics*, **88**.2:021002, 2016.

[7] K.-T. Chen. Integration of paths—a faithful representation of paths by non-commutative formal power series. *Trans. Amer. Math. Soc.*, **89**:395–407, 1958.

[8] P. Cheridito and B. Gersey. Computation of Conditional Expectations with Guarantees. *J. Sci. Comput.*, **95**.12:1–30, 2023.

[9] I. Chevyrev and T. Lyons. Characteristic functions of measures on geometric rough paths. *Ann. Probab.*, **44**.6:4049–4082, 2016.

[10] I. Chevyrev and H. Oberhauser. Signature moments to characterize laws of stochastic processes. *J. Mach. Learn. Res.*, **23**.176:1–42, 2022.

[11] Samuel N Cohen et al. Nowcasting with signature methods. *arXiv preprint arXiv:2305.10256*, 2023.

[12] J. B. Conway. *A Course in Functional Analysis.* Second Edition. Graduate Texts in Mathematics **96**, Springer, 1997.

[13] J. Elstrodt. *Maß- und Integrationstheorie.* Achte, erweiterte und aktualisierte Ausgabe, Springer Spektrum, 2018.

[14] P. K. Friz and N. B. Victoir. *Multidimensional Stochastic Processes as Rough Paths: Theory and Applications.* Cambridge Studies in Advanced Mathematics **120**, Cambridge University Press, 2010.

[15] R. Giles. A Generalization of the Strict Topology. *Trans. Amer. Math. Soc.*, **161**:467–474, 1971.

[16] B. Hambly and T. Lyons. Uniqueness for the signature of a path of bounded variation and the reduced path group. *Ann. of Math.*, **171**.1:109–167, 2010.

[17] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Second Edition, Springer Series in Statistics, 2009.

[18] S. Janson. *Gaussian Hilbert Spaces.* Number **129**. Cambridge University Press, 1997.

[19] O. Kallenberg. *Foundations of Modern Probability.* Springer, 2021.

[20] I. Karatzas and S. Shreve. *Brownian Motion and Stochastic Calculus*, volume 113. Springer Science & Business Media, 1998.

[21] A Kechris. *Classical Descriptive Set Theory.* Graduate Texts in Mathematics **156**, Springer, 1995.

[22] J.P. Klein and M.L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data.* Springer, 2003.

[23] K. Kuratowski. *Topology.* Volume 1. New edition, revised and augmented. Academic Press, New York 1966.

[24] F. Legoll and T. Lelievre. Effective Dynamics using Conditional Expectations. *Nonlinearity*, **23**.9:2131, 2010.

[25] D. Levin, T. Lyons, and H. Ni. Learning from the past, predicting the statistics for the future, learning an evolving system. *arXiv preprint arXiv:1309.0260*, 2013.

[26] A. Lewis. Option Valuation Under Stochastic Volatility. *Finance Press*, 2000.

[27] S. Liao, H. Ni, M. Sabate-Vidales, L. Szpruch, M. Wiese, and B. Xiao. Sig-Wasserstein GANs for Conditional Time Series Generation. *Math. Finance*, 2023.

[28] T. J. Lyons. Differential Equations Driven by Rough Signals. *Revista Matemática Iberoamericana*, **14**.2:215–310, 1998.

[29] T. J. Lyons, M. Caruana, and T. Lévy. *Differential Equations Driven by Rough Paths.* Springer, 2007.

[30] H. G. Matthies, E. Zander, B. V. Rosić, and A. Litvinenko. Parameter Estimation via Conditional Expectation: a Bayesian Inversion. *Advanced Modeling and Simulation in Engineering Sciences*, **3**.1:1–21, 2016.

[31] M.J. Neely, E. Modiano, and C.-P. Li. Fairness and Optimal Stochastic Control for Heterogeneous Networks. *IEEE/ACM Transactions On Networking*, **16**.2:396–409, 2008.

[32] M. Ottaviani and P.N. Sørensen. The Strategy of Professional Forecasting. *Journal of Financial Economics*, **81**.2:441–466, 2006.

[33] J. M. Robins, S. D. Mark, and W. K. Newey. Estimating Exposure Effects by Modelling the Expectation of Exposure Conditional on Confounders. *Biometrics*, pages 479–495, 1992.

[34] Peter M Robinson. Nonparametric Estimators for Time Series. *J. Time Series Anal.*, **4**.3:185–207, 1983.

[35] L. C. G. Rogers and D. Williams. *Diffusions, Markov Processes, and Martingales.* Volume 1, Cambridge University Press, 2000.

[36] F. S. Scalora. Abstract Martingale Convergence Theorems. *Pacific J. Math.*, **11**.4:347–374, 1961.

[37] M.J. Schervish. Theory of Statistics. *Springer Series in Statistics*, 1995.

[38] M. Shanahan. Talking About Large Language Models. *arXiv preprint arXiv:2212.03551*, 2022.

[39] S.E. Shreve. *Stochastic Calculus for Finance II: Continuous-Time Models*, volume **11**. Springer, 2004.

[40] R. H Shumway and D. S. Stoffer. An Approach to Time Series Smoothing and Forecasting Using the EM Algorithm. *J. Time Series Anal.*, **3**.4:253–264, 1982.

[41] P. F. Shustin and H. Avron. Some Perturbation Results for a Normalized Non-Orthogonal Joint Diagonalization Problem. *SIAM J. Matrix Anal. Appl.*, **43**.1:479–511, 2022.

[42] J. M. A. M. Van Neerven. Stochastic Evolution Equations.

[43] S. Wolfram. *What Is ChatGPT Doing... and Why Does It Work?* Stephen Wolfram, 2023.

[44] J. Yeh. *Martingales and Stochastic Analysis.* Series on Multivariate Analysis **1**, World Scientific Publishing, 1995.

[45] S. L. Zeger and B. Qaqish. Markov Regression Models for Time Series: A Quasi-Likelihood Approach. *Biometrics*, pages 1019–1031, 1988.

[46] W. X. Zhao et al. A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223*, 2023.