ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Seminar for Applied Mathematics SAM

Nonparametric Regression of Stochastic Processes via Signatures

A. Schell and R. Alaifari

Research Report No. 2023-45

December 2023
Latest revision: May 2024

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

Nonparametric Regression of Stochastic Processes via Signatures

Alexander Schell* Rima Alaifari†

Seminar for Applied Mathematics, ETH Zurich

31 May 2024

Abstract

Nonparametric regression of stochastic processes estimates statistical relationships between multidimensional, time-dependent data without relying on specific parametric assumptions. We propose a novel approach to this classical estimation problem by using the signature transform from rough path theory to encode the information of a stochastic process into a sequence of iterated integrals, capturing its statistical properties in a time-global and hierarchical manner. Viewing statistical regression as an operator learning problem, this signature-based discretisation allows us to characterise the conditional statistical dependence of a stochastic process on another stochastic process as the solution to a convex semi-infinite linear least squares problem. This result is based on a functional monotone class argument involving the bounded signature of the conditioning process and allows for the efficient and provably consistent nonparametric estimation of regression functions and conditional distributions for very general classes of jointly distributed stochastic processes as solutions to convex optimisation problems. The structural insights of this approach are summarised in two universally consistent regression estimators that are computable with practical algorithms and supported by broad theoretical guarantees.

Keywords: stochastic processes, nonparametric and functional regression, conditional mean embedding, conditional expectation, signature features, kernel ridge regression, function approximation

1 Introduction

Modelling and inferring meaningful patterns and relationships within complex, high-dimensional data is a central challenge in modern machine learning and statistics. At the core of this challenge lies the task of capturing statistical dependencies between data sets through conditional distributions and the structured operationalisation of statistical conditioning. These are fundamental aspects of probabilistic modelling that are crucial for a growing range of inference techniques and applications. One of the most important paradigms for this is nonparametric regression, which is predicated on the idea that the relationships between variables can be captured with minimal assumptions about their functional form. Unlike parametric methods, which specify a fixed structure for the relationship (e.g., linear or polynomial), nonparametric approaches allow the data to speak for itself, revealing the underlying patterns and relationships more flexibly.

Conceptually, nonparametric regression involves estimating a regression function that describes the conditional expectation of a response variable given one or more predictor variables. The challenge intensifies when the predictors or responses are observed as part of a stochastic process, since

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

one must then account for temporal dependencies and potential non-stationarities inherent in the data. Applications of such nonparametric regression methods for stochastic processes are extensive and profound. In finance and econometrics, these methods can be used to model asset prices, interest rates, or volatility, capturing subtle dependencies and predicting future trends [54]. They also permeate risk assessments in financial markets and portfolios [2], or stochastic filtering and the optimisation of control systems in engineering [3, 38]. In environmental science, they help in understanding climate patterns, pollution levels, or ecological dynamics, accommodating complex interactions and temporal variations. In biomedical engineering, they are employed to analyse dependencies between physiological signals, such as heart rate variability or brain activity, where the underlying biological processes are inherently stochastic and nonlinear. Stochastic process regression is also applied in computer vision [1] and molecular dynamics [30], and it is central to survival analysis [27] and causal models [44], time series analysis and forecasting [39, 45, 55], Bayesian inversion and inference [37, 50] and statistical machine learning [21] broadly, to name just a few classical examples. Most recently, nonparametric stochastic process regression has emerged as the key concept behind large language models [62], which are essentially high-dimensional statistical regression models that approximate conditional distributions on sequential data [53, 59].

A particular challenge in computing conditional expectations and their derived statistics, present in most of the above examples and especially in sequential or language-based machine learning, is to efficiently account for potential time dependencies in the conditioning variable, i.e. the conditioning on complex multidimensional 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. However, these classical 'convenience' assumptions have clear limitations (e.g. [6, 32]), making it worthwhile to revisit the general problem of time-dependent regression and conditioning with modern tools from stochastic analysis. The present work aims to contribute to this endeavour.

Specifically, this paper addresses the following central inference questions, which the above examples suggest are of significant practical relevance: Given two multidimensional stochastic processes X and Y in discrete- or continuous time and some random vector Z in Euclidean space,

$$\textit{how can the statistics} \quad f \mapsto \mathbb{E}[f(Y) \,|\, X] \quad \textit{and} \quad \mathbb{E}[Z \,|\, X] \quad \textit{be efficiently estimated?} \tag{1}$$

(The conditional law $\mathbb{P}(Y \in \cdot \,|\, X)$ is included by letting f run over the indicator functions on Y.) The approximation of such conditional expectations defines the problem of statistical regression. This has been well-addressed for time-independent X or if the temporality of $X = (X_t)$ conforms to certain parametric assumptions (e.g. [10, 5, 61] and the references therein). However, to the best of our knowledge, there are currently no rigorous nonparametric solutions to (1) for the case of 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 bounded (or "robust") signatures recently introduced in [12], to structure the statistical information given by the predictor process X: A bounded signature, ϕ, is an algebraically structured and bounded Hilbert-valued coordinate map over the space of sufficiently continuous paths, which includes the sample realisations of X. Using a functional monotone class argument, we show that for a linear subset $\mathfrak{L}$ of the Hilbert co-domain $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ of ϕ, the coordinate functions $\{\langle \ell, \phi(X) \rangle \mid \ell \in \mathfrak{L}\}$ form an L^2-dense subspace of all square-integrable X-measurable random variables. As a basic step toward answering (1), this then implies the variational characterisation

$$\mathbb{E}[Z \,|\, X] \;=\; \lim_{k \to \infty} \Big(\big\langle \alpha_{1k}, \phi(X) \big\rangle, \cdots, \big\langle \alpha_{mk}, \phi(X) \big\rangle \Big), \tag{2}$$

with m the dimension of $Z \equiv (Z^1, \cdots, Z^m)$ and for any minimizing sequence $(\alpha_{1k}, \ldots, \alpha_{mk})_k$ of

$$\inf_{(\tilde{\alpha}_1,\ldots,\tilde{\alpha}_m) \in \mathfrak{L}^m} \sum_{i=1}^{m} \mathbb{E}\big|Z^i - \langle \tilde{\alpha}_i, \phi(X)\rangle\big|^2. \tag{3}$$

The parameter optimization (3) that underlies the above approximation is notably convex, and the convergence (2) holds at least in L^2 and almost surely if the sequence (α_{ik}) is fast enough. A similar characterisation can be found for the conditional expectation given X of the bounded signatures of Y itself, leading to the variational identities (convergent in L^2, and under conditions almost surely)

$$\mathbb{E}[f(Y)\,|\,X] \;=\; \lim_{l\to\infty} \lim_{k\to\infty} \big\langle \ell_l^f\,, \psi_{\alpha_k}(X)\big\rangle \tag{4}$$

for any measurable function f of Y such that $f(Y)$ square-integrable. In this case, $(\psi_{\alpha_k})_k$ is an L^2-dense adaptive system of ϕ-derived model functions on sample paths that conceptually resembles (and effectively replaces) an 'algebraically structured and convexly parametrisable neural network', and $(\ell_l^f)_l$ is a dualised signature-encoding of the argument function f. Optimally parametrised instances of (ψ_{α_k}) and the function encodings (ℓ_l^f) can both be computed explicitly as solutions to two separate, well-structured and efficiently approximable convex optimisation problems.

In addition to being fully convex-optimisable, the representations (2) and (4) hold nonparametrically for any response-transforming nonlinearity f without needing any assumptions on the relationship between the marginals in (X, Z) or (X, Y), such as continuity or specific statistical dependence structures. The flexible and comparatively simple algorithmic premise behind these identities also makes the regression statistics in (1) amenable to nonparametric statistical estimation within the well-established empirical framework of kernel ridge regression. Consequently, this approach provides a theoretically and practically attractive solution to the inference problem (1).

This paper is structured as follows. Section 2 reviews the fundamental concepts behind nonparametric regression analysis and statistical conditioning and explains their use for modelling and analysing dependencies between statistically associated datasets (Sections 2.1 and 2.2). It then sets the stage for (1) by reformulating this question as an equivalent operator learning problem using the perspective of conditional mean embedding (Section 2.3). Leaving the general setting behind, Section 3 delves into the specifics of nonparametric regression of time-dependent data, covering structural basics for the associated regression spaces and time-dependent random variables (Section 3.1). The signature transform is introduced as a key tool for representing such time-dependent data (Section 3.2), complemented by a brief discussion of its basic properties and how it can serve as a bounded global coordinate map on path spaces (Section 3.3). A derivation of the variational identities (2) and (4) is provided in Section 4, which presents the main theoretical contributions of the paper. We show how the bounded signature can be used to discretise and exhaust the information of the conditioning process (Section 4.1), and utilise this discretisation to characterise the conditional expectations in (1) as solutions of convex semi-infinite linear least squares problems (Sections 4.2 and 4.3). These characterisations are then promoted to practical estimators for the nonparametric regression statistics (1) in Section 5, which builds on the theoretical insights of the previous sections to achieve an efficient data-based approximation of the proposed representations (4) and (2). Upon embedding our signature-based regression architectures in the context of vector-valued reproducing kernel Hilbert spaces and providing basic support theory for a subsequent analysis of statistical convergence properties (Section 5.1 and 5.2), we present explicitly computable nonparametric regression estimators for (1) (Sections 5.3 and 5.4) and provide an analysis of their statistical approximation properties, including convergence rates and error bounds (Section 5.5). Numerical experiments and practical example applications of our method are to be included in the forhtcoming arXiv version of this paper. Most technical proofs are provided in the Appendix.

Closing with a note on existing literature, we remark that prior approximations in a manner similar to (2) were first explored in [31, 34], though the respective aspects of these works are mostly empirical and based on rather strict assumptions on (X, Z). In a spirit related to ours but with no immediate mathematical connection, [13] 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 made independently of us (and by other mathematical means) in the recent preprint [4], where this idea is applied to different consequences in the realm of optimal stopping.

2 Perspectives on Regression and Statistical Conditioning

This section provides a brief review of probabilistic regression and statistical conditioning, introducing basic concepts (Sections 2.1 and 2.2) as well as more specialised mathematical perspectives (Section 2.3) to structure and operationalise the underlying statistical theory. The aim is to lay bare the essential probabilistic structure behind regression-based statistical inference and learning on Polish spaces, so that, in subsequent sections, we can seamlessly extend this foundational theory in the particular setting of nonparametric regression for time-dependent multidimensional data.

Specifically, Section 2.1 motivates the framework of statistical regression analysis as a principled approach to model and analyse complex, non-functional statistical dependencies between two sets of data residing in potentially infinite-dimensional spaces. The classical idea is to view these data as samples from a pair of jointly distributed random variables and decompose their joint probability distribution into a family of conditional distributions with respect to one of their marginals. This is formalised through the well-known concept of disintegration, which Section 2.2 briefly recalls and contextualises. With these basics in hand, Section 2.3 draws on the perspective of conditional mean embedding to reformulate the problem of nonparametric (and, in general, nonlinear) regression as an equivalent, yet more tangible problem of approximating, from finite data, a bounded linear 'regression' operator (Proposition 2.4 and Definition 2.6; also Lemma 2.9). We show that this problem can be solved under a 'sufficiently nonlinear' Hilbert-valued coordinatisation (15) over the space of all response data (Proposition 2.4), and illustrate how this approach yields a comprehensive statistical framework for stochastic process regression and in the context of large language models.

2.1 Regression Analysis: Modelling Statistical Dependencies in Data

Extracting and analysing statistical patterns and dependencies from complex, high-dimensional data through conditional distributions and statistical conditioning is a fundamental problem central various inference techniques and applications.

One of the most popular approaches to this challenge is regression analysis, which aims to identify the relationship between two given data sets (x_i) and (y_i). Here, each x_i represents an input and each y_i represents a corresponding output, and these data points reside in potentially infinite-dimensional spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively. Classical regression analysis, and most physical models alike, aims to extract from the pairs (x_i, y_i) an essentially functional relationship of the form:

$$\mathfrak{R}^f = \big\{(x, f(x)) \,\big|\, x \in \mathcal{X}\big\} \quad \text{or, more broadly,} \quad \mathfrak{R} = \{(x, y) \in \mathcal{X} \times \mathcal{Y} \mid y \approx f(x)\}, \tag{5}$$

assuming the existence of a function $f : \mathcal{X} \to \mathcal{Y}$ that maps inputs $\{x\}$ to outputs $\{y\}$. This f then encapsulates the 'true' link between the data (x_i) and (y_i). It is usually approximated by a proxy $f_{\theta_\star}$ selected from a predefined family of model functions (f_θ) to best fit the observed pairs (x_i, y_i).

However, many real-world scenarios defy this simplified 'functional' assumption, presenting cases where an input x in $\mathcal{X}$ does not correspond to a singular deterministic output $f(x)$ in $\mathcal{Y}$, but to a

range $\mathfrak{R}_x^{\sim}$ of possible outcomes associated with x. For instance, in machine translation, sentences x from a source language $\mathcal{X}$ are to be translated into sentences y in a target language $\mathcal{Y}$. This defines the complex relation

$$\mathfrak{R}^{\sim} := \big\{(x,y) \in \mathcal{X} \times \mathcal{Y} \;\big|\; y \text{ is a valid translation of } x\big\}, \tag{6}$$

which typically extends beyond a simple functional relation ($y \approx f(x)$) since a given sentence x in $\mathcal{X}$ can translate into multiple valid sentences y in $\mathcal{Y}$. Such a non-functional, multi-valued relationship between x and y is termed an *associative relation*, or an *association*, between x and y.

To handle these complexities, classical regression analysis can be extended into a rigorous probabilistic framework. Instead of describing associations through a deterministic function from $\mathcal{X}$ to $\mathcal{Y}$, a more appropriate approach is to use a (set-valued) map from $\mathcal{X}$ to $2^{\mathcal{Y}}$, sending x to $\mathfrak{R}_x^{\sim} := \{y \in \mathcal{Y} \mid (x,y) \in \mathfrak{R}\}$. Further complexity is captured by assigning to each $x \in \mathcal{X}$ a probability measure μ_x over $\mathfrak{R}_x^{\sim}$, reflecting the likelihood of different outcomes associated with x. One way to implement this is by defining a probability measure μ on $\mathcal{X} \times \mathcal{Y}$, say $\mu = \mathbb{P}_{(X,Y)}$ for an $\mathcal{X} \times \mathcal{Y}$-valued random variable (X,Y) that models the data. (The marginal variables X and Y are called regressor and regressand, respectively.) The desired plausibility measures $(\mu_x)_{x \in \mathcal{X}}$ are then obtained as the conditional probabilities of Y given $X = x$. This probabilistic approach provides a well-known framework to describe associative relationships between data (x) and (y) through an analysis of the statistical dependency between the random variables X and Y that model the data.

The remainder of Section 2 reviews the necessary mathematics to develop this perspective into a coherent statistical theory. The focus of our discussion will be on how to consistently approximate the probabilistic model $(\mu_x)_{x \in \mathcal{X}}$ of an associative relationship between data from a finite number of samples. We henceforth assume that $\mathcal{X}$ and $\mathcal{Y}$ are Polish spaces, unless otherwise stated.

2.2 Capturing Statistical Dependencies with Probability Kernels

As discussed above, the main idea behind the probabilistic modelling of an associative relation in $\mathcal{X} \times \mathcal{Y}$ (such as (6)) is to reframe this relation as a functional relation within $\mathcal{X} \times \mathcal{M}_1(\mathcal{Y})$:

> By assuming that the statistically dependent ('associated') data (x_i) and (y_i) are sampled from a joint distribution μ on $\mathcal{X} \times \mathcal{Y}$, say $(x_i, y_i) \sim \mu$, we aim to *decompose* μ into a measure-valued function $x \mapsto \mu_x$ from $\mathcal{X}$ to $\mathcal{M}_1(\mathcal{Y})$, such that for each $x \in \mathcal{X}$, the measure $\mu_x \in \mathcal{M}_1(\mathcal{Y})$ represents the conditional law of data in $\mathcal{Y}$ given x.

This concept is rigorously defined through the classical notion of a probability kernel. This is a map $\kappa : \mathcal{X} \times \mathcal{B}(\mathcal{Y}) \to [0,1]$ with $(\kappa(x,\,\cdot\,) \mid x \in \mathcal{X}) \subseteq \mathcal{M}_1(\mathcal{Y})$ such that for each set $B \in \mathcal{B}(\mathcal{Y})$,

$$\mathcal{X} \ni x \;\longmapsto\; \kappa(x,B) =: \kappa_x(B) \quad \text{is} \quad \big(\mathcal{B}(\mathcal{X}), \mathcal{B}([0,1])\big)\text{-measurable.}^{1}$$

Following this, recall that every such kernel κ can be 'fused' with any $\upsilon \in \mathcal{M}_1(\mathcal{X})$ via the *coupling*

$$\kappa \otimes \upsilon \;:\; \mathcal{B}(\mathcal{X} \times \mathcal{Y}) \ni A \;\longmapsto\; \int_{\mathcal{X}} \left[\int_{\mathcal{Y}} \mathbb{1}_A(x,y)\, \kappa_x(\mathrm{d}y) \right] \upsilon(\mathrm{d}x) \;\in\; [0,1], \tag{7}$$

and the resulting map (7) is then again a Borel probability measure on $\mathcal{X} \times \mathcal{Y}$. Conversely, any given measure $\mu \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ can be 'divided' by each of its marginals, and the resulting 'quotient' is essentially unique. This is formalised by the concept of *disintegration*, which asserts that there is

$$\text{a 'unique' probability kernel} \quad \mu_{\mathcal{Y}|\mathcal{X}} \equiv \big(\mu_{\mathcal{Y}|x} \mid x \in \mathcal{X}\big) \quad \text{such that} \quad \mu \,=\, \mu_{\mathcal{Y}|\mathcal{X}} \otimes \mu_{\mathcal{X}}\,, \tag{8}$$

where 'unique' means 'unique up to inequality on a $\mu_{\mathcal{X}}$-nullset': for any two kernels $\kappa, \tilde{\kappa} \subseteq \mathcal{M}_1(\mathcal{Y})$, the identity $\kappa \otimes \mu_{\mathcal{X}} = \tilde{\kappa} \otimes \mu_{\mathcal{X}}$ implies that $\kappa_x = \tilde{\kappa}_x$ $\mu_{\mathcal{X}}$-almost everywhere (see e.g. [23, Chap. 8]).

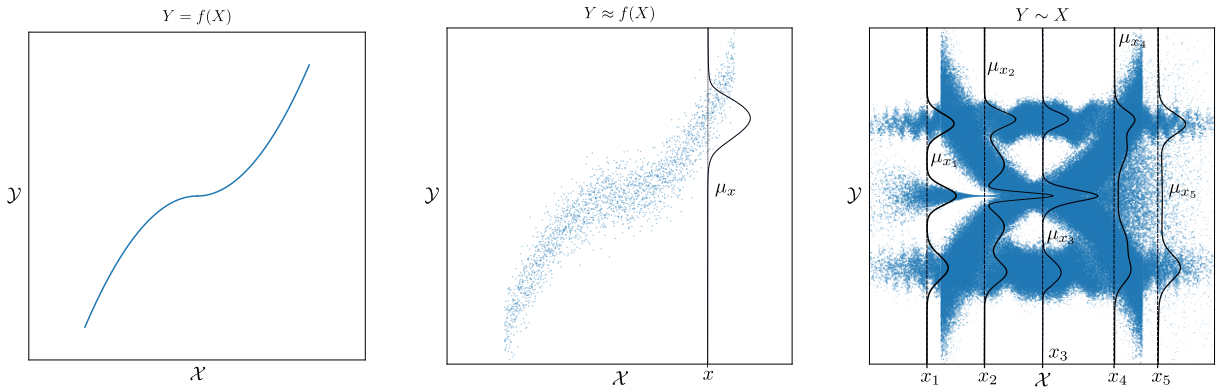

Figure 1: *Visualisation of increasingly non-functional relations and their probabilistic description. The left panel shows a functional relation $f : x \mapsto y$, as in (5). The middle panel depicts a (slightly non-functional) associative relation where an input x maps to a whole set $\mathfrak{R}_x$ of possible outputs, together with an associated probability measure μ_x capturing the likelihood of different outcomes within $\mathfrak{R}_x$. The right panel shows a fully associative relation where X and Y exhibit complex dependencies; it also demonstrates the concept of disintegrations (8), showing the decomposition of the joint distribution $\mu = \mathbb{P}_{(X,Y)}$ into the conditional distributions $\mu_{\mathcal{Y}|\mathcal{X}} \equiv (\mu_x)_{x\in\mathcal{X}}$ wrt. the marginal $\mu_{\mathcal{X}} = \mathbb{P}_X$. Altogether, these panels illustrate the transition from classical regression to more complex probabilistic models and the role of disintegrations (12) in capturing statistical dependencies.*

Consequently, for a fixed $\mathcal{X}$-marginal $\xi \in \mathcal{M}_1(\mathcal{X})$, the assignment $\mu \mapsto \mu_{\mathcal{Y}|\mathcal{X}}$ in (8) defines a bijection

$$\mathfrak{c}_\xi \;:\; \mathcal{M}_1^\xi(\mathcal{X}\times\mathcal{Y}) \longrightarrow L^0\big(\xi;\mathcal{M}_1(\mathcal{Y})\big), \quad \mu \;\mapsto\; \mu_{\mathcal{Y}|\mathcal{X}}, \tag{9}$$

with inverse $\mathfrak{c}_\xi^{-1}(\,\cdot\,) = (\,\cdot\,)\otimes\xi$. This *conditioning* map (9) provides the desired 'functionalisation' of any type of statistical dependence between the marginals of $\{(x_i, y_i)\}$. The map (9) thus captures every kind of associative relation on $\mathcal{X}\times\mathcal{Y}$ by interpolating between two statistical extreme cases:

Example 2.1 (Functional Dependence and Independence)**.** Given marginals $\xi \in \mathcal{M}_1(\mathcal{X})$ and $\mu_{\mathcal{Y}} \in \mathcal{M}_1(\mathcal{Y})$—read as the distributions of (x_i) and (y_i), resp.—and a function $f : \mathcal{X} \to \mathcal{Y}$, consider

$$\mu^f_{\mathcal{Y}|\mathcal{X}} \coloneqq (\delta_{f(x)} \mid x \in \mathcal{X}) \qquad \text{and} \qquad \mu^\perp_{\mathcal{Y}|\mathcal{X}} \coloneqq (\mu_{\mathcal{Y}} \mid x \in \mathcal{X}).$$

These measure-valued functions (in fact kernels, i.e. in $L^0\big(\xi;\mathcal{M}_1(\mathcal{Y})\big)$, if f is Borel-measurable) identify associations $\mu_f \coloneqq \mu^f_{\mathcal{Y}|\mathcal{X}}\otimes\xi$ $(=\mathfrak{c}_\xi^{-1}(\mu^f_{\mathcal{Y}|\mathcal{X}}))$ and $\mu_\perp \coloneqq \mu^\perp_{\mathcal{Y}|\mathcal{X}}\otimes\xi = \mu_{\mathcal{Y}}\otimes\xi$ $(=\mathfrak{c}_\xi^{-1}(\mu^\perp_{\mathcal{Y}|\mathcal{X}}))$ that characterize functional dependence (5) and statistical independence ($(x_i)\perp(y_i)$), respectively. ◆

The concept of a conditioning map (9) allows us to express the task of learning associative relations between data (from finitely many data samples) as follows:

$$\textit{For}\;\; \mu \in \mathcal{M}_1^\xi(\mathcal{X}\times\mathcal{Y}),\;\; \textit{approximate the kernel}\;\; \mathfrak{c}_\xi(\mu)\;\; \textit{from samples of}\;\; \mu. \tag{10}$$

Before we transition from the informal description (10) (the term 'approximate' is made precise later) to a precise and actionable problem formulation, let us first establish some basic terminology.

Definition 2.2. For $\mathcal{X},\mathcal{Y}$ Polish spaces and $\xi \in \mathcal{M}_1(\mathcal{X})$, the disintegration $\mathfrak{c}_\xi(\mu)$ of any law $\mu \in \mathcal{M}_1^\xi(\mathcal{X}\times\mathcal{Y})$ is termed the *conditional dependence of μ wrt. the source law ξ*.

It is often convenient, not least for notational reasons, to view a measure $\mu \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ as the law of some $(\mathcal{X} \times \mathcal{Y})$-valued random variable. That is, to suppose that

$$\mu = \mathbb{P}_{(X,Y)} \qquad \text{for some} \qquad X : \Omega \to \mathcal{X} \quad \text{and} \quad Y : \Omega \to \mathcal{Y}, \tag{11}$$

where both X and Y are some Borel-measurable maps over a joint probability space $(\Omega, \mathscr{F}, \mathbb{P})$; the data $\{(x_i, y_i)\}$ are then sample realisations of (X, Y). In common regression terminology, Y is called the 'regressand' and X is called the 'regressor'. In this setting, the marginals of μ are $\mu_{\mathcal{X}} = \mathbb{P}_X$ and $\mu_{\mathcal{Y}} = \mathbb{P}_Y$, and it is customary to denote the conditional kernel $\mu_{\mathcal{Y}|\mathcal{X}}$ in (8) as

$$\mathfrak{c}_X(\mu) \equiv \big(\mu_{\mathcal{Y}|x}(\,\cdot\,) \,:\, x \in \mathcal{X}\big) =: \big(\mathbb{P}(Y \in (\,\cdot\,) \mid X = x) \,:\, x \in \mathcal{X}\big). \tag{12}$$

A solution to (10) then provides a data-based approximation ('estimator') of the statistics

$$\begin{aligned} &\mathbb{P}(Y \in A \mid X = x) = \int_A \mathrm{d}\mu_{\mathcal{Y}|x}, \quad \mathbb{P}(X \in A, Y \in B) = \int_A \int_B \mathrm{d}\mu_{\mathcal{Y}|x}\, \mathbb{P}_X(\mathrm{d}x) \quad \text{and} \\ &\mathbb{E}[f(Y) \mid X] = \int_{\mathcal{Y}} f \,\mathrm{d}\mu_{\mathcal{Y}|X}, \qquad \text{for any} \quad A \times B \in \mathcal{B}(\mathcal{X} \times \mathcal{Y}) \;\text{ and }\; f \in \mathcal{L}^2(\mathbb{P}_Y), \end{aligned} \tag{13}$$

where $\mu_{\mathcal{Y}|X} \coloneqq \mathfrak{c}_\xi(\mu) \circ X$. These are of central importance to various applications.

Remark 2.3. Any conditional dependence (CD) can be lifted to the (conditional) law of a canonical $(\mathcal{X} \times \mathcal{Y})$-valued random variable, and the converse is also true, giving a one-to-one correspondence between CDs and paired random variables, see Remark B.1. Thus, the 'lifting assumption' (11) entails no loss of generality and is, in fact, not an assumption but rather a provable statement. ◆

2.3 Nonparametric Regression as an Operator Learning Problem

Given a law $\mu \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ with marginal $\xi \coloneqq \mu_{\mathcal{X}}$, the main challenge we want to address is how to efficiently approximate the conditional dependence $\mathfrak{c}_\xi(\mu)$ from a finite number of μ-drawn samples. In fact, we will tackle the task (10) in the setting of time-dependent data, where $\mathcal{X}$ and $\mathcal{Y}$ are infinite-dimensional function spaces consisting of vector-valued paths (Sect. 3). This sequential setting poses several statistical and analytical challenges, not least because the resulting relation spaces $\mathcal{X} \times \mathcal{Y}$ are generally not locally compact. Our approach to handle this is to work with sufficiently nonlinear 'coordinate maps' on $\mathcal{X}$ and $\mathcal{Y}$ that ultimately allow us to estimate (13) with bespoke approximation architectures grounded in rigorous statistical theory (see Sections 4 and 5).

But first, we must ask how, and in what precise sense, we can approximate a conditional dependence (8)—an uncountable family of measures—from finite data. An established[2] approach is to use functional analysis and view such families as explicitly representable bounded linear operators.

Proposition 2.4 (Conditional Mean Embedding)**.** *For any law $\mu \in \mathcal{M}_1^\xi(\mathcal{X} \times \mathcal{Y})$, the conditional dependence $\mathfrak{c}_\xi(\mu) \equiv (\mu_x)_{x \in \mathcal{X}} \subseteq \mathcal{M}_1(\mathcal{Y})$ of μ wrt. ξ can be identified with a bounded linear operator*

$$\mathfrak{c}_\xi(\mu) \,:\, L^2(\mu_{\mathcal{Y}}) \longrightarrow L^2(\xi), \quad f \longmapsto \mu_{\cdot}(f) \coloneqq \left[x \mapsto \int_{\mathcal{Y}} f \,\mathrm{d}\mu_x\right], \tag{14}$$

Moreover, given a separable Hilbert space $\big(\mathcal{H}_{\mathcal{Y}}, \langle \cdot, \cdot \rangle\big)$ together with a Borel-measurable map

$$q : \mathcal{Y} \to \mathcal{H}_{\mathcal{Y}} \quad \textit{such that} \quad \big\{\langle \ell, q(\cdot)\rangle \mid \ell \in \mathcal{H}_{\mathcal{Y}}\big\} \;\textit{is } L^2\textit{-dense in } L^2(\mu_{\mathcal{Y}}), \tag{15}$$

[2] See, for instance, [41, 33] and the references therein.

then the (so-called) regression operator (14) *and, thus, the conditional dependence* $\mathfrak{c}_\xi(\mu)$ *itself, can be identified with the function*

$$\mu_{\cdot}(q) \in L^2(\xi;\mathcal{H}_\mathcal{Y}) \quad \textit{given by} \quad x \mapsto \mu_x(q) \coloneqq \int_\mathcal{Y} q(y)\,\mu_x(\mathrm{d}y). \tag{16}$$

Specifically then, for each $f \in L^2(\mu_\mathcal{Y})$ *there is a sequence* $(\ell_j^{(f)})_j \subset \mathcal{H}_\mathcal{Y}$ *such that*

$$\mathfrak{c}_\xi(\mu)(f) \;=\; \lim_{j\to\infty} \big\langle \ell_j^{(f)}, \mu_{\cdot}(q)\big\rangle \quad \textit{in} \quad L^2(\xi). \tag{17}$$

(*For functions* f *such that* $f = \langle \ell, q(\cdot)\rangle$ *for some* $\ell \in \mathcal{H}_\mathcal{Y}$*, we can choose* $\ell_j^{(f)} \coloneqq \ell$ *for each* j.)

Proof. Mostly straightforward consequences of the definitions, but see Appendix B.1.2. □

Remark 2.5 (Parametrizing Conditional Dependencies)**.** The representation (17) based on (15) can be seen as an 'infinite-dimensional parametrization' of the regression operator (14) representing a conditional dependence $\mathfrak{c}_\xi(\mu)$; see also Remark B.2. The parameter domain for this representation can be chosen as the L^2-closure of a separable Hilbert space. For conditional dependencies between time-dependent data, this domain has a highly-structured and conveniently explicit form; see Proposition 4.10 and Lemma 3.12. These structural benefits, as shown in Section 5, facilitate translating the operator representation (17) into an efficient scheme for estimating $\mathfrak{c}_\xi(\mu)$. ◆

The functional analytic viewpoint of Proposition 2.4—specifically, the one-to-one correspondence between (9) and (14)—facilitates a rigorous and straightforward formulation of (10) in terms of operator learning. For any fixed pair $\mathcal{X},\mathcal{Y}$ of Polish spaces, denote $\mathcal{Z} \coloneqq \mathcal{X}\times\mathcal{Y}$ and let $\mathscr{Z} \coloneqq \mathcal{Z}^\mathbb{N}$.

Definition 2.6 (Learning CDs)**.** Let $\mu \in \mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$. A sequence of operator-valued maps

$$\hat{T}_n \,:\, \mathscr{Z} \longrightarrow \big[\mathcal{L}^2(\mu_\mathcal{Y}) \to \mathcal{L}^2(\mu_\mathcal{X})\big] \qquad (n\in\mathbb{N}) \tag{18}$$

will be called *a consistent estimator of* $\mathfrak{c}_{\mu_\mathcal{X}}(\mu)$ if, for any iid samples $Z_1, Z_2, \ldots \sim \mu =: \mathbb{P}$, we have

$$\hat{T}_n\big((Z_j)\big) \xrightarrow[n\to\infty]{\mathbb{P}} \mathfrak{c}_{\mu_\mathcal{X}}(\mu) \quad \text{strongly,} \quad \text{that is:} \quad \lim_{n\to\infty} \mathbb{P}\Big(\big\|\mu_{\cdot}(f) - \hat{T}_n\big((Z_j)\big)(f)\big\|_{L^2(\mu_\mathcal{X})} \geq \varepsilon\Big) = 0 \tag{19}$$

for each $f \in \mathcal{L}^2(\mu_\mathcal{Y})$, which is required to hold for any $\varepsilon > 0$. We call *universally consistent* a sequence $(\hat{T}_n)$ for which both (18) and (19) hold for every measure $\mu \in \mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$.

Remark 2.7 (Stochastic Process Regression)**.** The main goal of this paper is to propose a universally consistent estimator of conditional dependencies on spaces $\mathcal{X},\mathcal{Y}$ of multivariate time-dependent data. In other words (see Remark 2.3), we aim to estimate the conditional dependencies between coupled multidimensional (discrete- or continuous-time) stochastic processes X and Y. In this context, we recognize in (14) the familiar expressions

$$\mu_{\cdot}(f) \,:\, x \mapsto \mu_x(f) \equiv \mathbb{E}[f(Y)\,|\,X=x] \quad \text{and} \quad \mu_X(f) \coloneqq \mu_{\cdot}(f)\circ X \equiv \mathbb{E}[f(Y)\,|\,X] \tag{20}$$

(recall (13)), or in other words, $\mu_{\cdot}(f)$ is precisely the *regression function of* $f(Y)$ *on* $\mathcal{X}$. In particular,

$$\mathbb{P}(Y\in(\cdot)\,|\,X=x) \equiv \mu_x(\mathbb{1}_{(\cdot)}) \,:\, \mathcal{B}(\mathcal{Y}) \ni A \longmapsto \mu_x(\mathbb{1}_A)\,\big(=\mu_x(A)\big) \in [0,1] \tag{21}$$

is the conditional distribution of Y given $X=x$, cf. (12), where $\mathbb{P}(Y\in(\cdot)\,|\,X) \coloneqq \mu_X(\mathbb{1}_{(\cdot)})$ as usual. If the space $\mathcal{Y}$ of possible outcomes is finite, then we have the explicit embedding

$$\mathbb{P}(Y\in A\,|\,X) \;=\; \sum_{y\in\mathcal{Y}} \mathbb{1}_A(y)\mu_X(\{y\}) \overset{(15)}{=} \lim_{j\to\infty} \textstyle\sum_{y\in\mathcal{A}} \big\langle \ell_j^{(y)}, \mathbb{E}[q(Y)\,|\,X]\big\rangle \qquad \big(A\in\mathcal{B}(\mathcal{Y})\big) \tag{22}$$

of the conditional law $\mathbb{P}(Y \in \cdot\,|X)$, where $\{(\ell_j^{(y)}) \,|\, y \in \mathcal{Y}\} \subset \mathcal{H}_\mathcal{Y}$ is some (deterministic) family of extractors obtainable as minimizing sequences of the regression problem $\inf\{\|\mathbb{1}_{\{y\}} - \langle \ell, q\rangle\|^2_{L^2(\mathbb{P}_Y)} \mid \ell \in \mathcal{H}_\mathcal{Y}\}$ (Section 4.3). A consistent estimator $(\hat{T}_n)$, as in (18) and (19), then yields an approximation

$$\lim_{n\to\infty} \mathbb{P}\big(\sup_{A\in\mathcal{B}(\mathcal{Y})} \big|\mathbb{P}(Y \in A \,|\, X) - \hat{T}_n(A)\big| \geq \varepsilon\big) = 0 \tag{23}$$

for any given $\varepsilon > 0$, where $\hat{T}_n(A) := \hat{T}_n(\mathbb{1}_A)$ (see Section 5.5). The representation (17) (if it exists) suggests a general blueprint for such estimators $(\hat{T}_n)$, namely a 'composite architecture' of the form

$$\hat{T}_n(f) \;=\; \big\langle \hat{\ell}_n^{(f)}, \hat{\Xi}_n^q \big\rangle \tag{24}$$

for each $f \in \mathcal{L}^2(\mathbb{P}_Y)$, where $\hat{\ell}_n^{(f)}$ and $\hat{\Xi}_n^q$ are estimators of the components $(\ell_n^{(f)})$ and $\mu_{\cdot}(q) \equiv \mathbb{E}[q(Y) \,|\, X = \cdot\,]$ from (17), respectively. This paper presents two universally consistent instances of this architecture and analyses their statistical approximation properties (19) (see Sects. 5.3–5.5). ◆

A prominent recent application of this regression framework for stochastic processes is found in computational linguistics:

Example 2.8 (Large Language Models)**.** A particular use case of the above regression framework is the efficient representation and estimation of conditional distributions (21) of a [random vector or] stochastic process Y given a stochastic process X, as outlined in (22) and (23). This is the central task of Large Language Models (LLMs) such as Generative Pre-trained Transformers (GPT), whose goal it is to predict—that is, to sample from an estimated conditional distribution—a contextually appropriate response $y \sim Y$ (potential text completions) given a sequential input $x \sim X$ (supplied text prompts). In this context, $\mathcal{X}$ is the space of all possible inputs, such as appropriately vectorized text prompts of a given length, and $\mathcal{Y}$ is the (finite, limited by the vocabulary size) set of all possible token sequences of a given length that make up the model's potential responses. An LLM learns from data a conditional dependence $\mathfrak{c}_\xi(\mu)$ as in (12), for μ the joint distribution of (X, Y) and ξ the law of X, and this $\mathfrak{c}_\xi(\mu)$ captures the complex statistical relationship between input sequences (X) and potential outcomes (Y). This learning is operationalised (as expressed by (19) and (23)) through an estimation of the conditional distribution $\mathbb{P}(Y \in (\cdot)\,|\,X)$, and in practice translates to computing the (empirical) likelihood of different possible model responses given a particular user input. (Given a prompt, the model calculates the probability of various next tokens based on the learned conditional distribution.) The necessary computations are performed by estimators $\hat{T}_n$ of the general form (18), which in contemporary models are typically implemented as sequence-processing neural networks with (empirically fine-tuned) architectures like transformers. These estimators serve to translate sampled text data into approximations of the conditional mean embeddings (14) and (17) (or other suitable representations of the probability kernel $\mathfrak{c}_\xi(\mu)$) that functionally express the associative relationship between X and Y. This yields a statistical approximation of the intricate relationship between input and output text data, underpinning the generative capabilities of LLMs and providing us with probabilistically founded function approximators capable of emulating the nuances and complexities of natural language. Applied to this context, the probabilistic concepts and novel statistical estimation methods in this paper may bring practical improvements and contribute to our understanding of the mathematical foundations of LLMs and statistical language models at large. For more information on the statistics of LLMs, see [**?**] and the references therein. ◆

A mathematically insightful interpretation of the regression operator $\mathfrak{c}_\xi(\mu)$ in (14) is to view it as providing, in an L^2-variational sense, the optimal transfer of information from $\mathcal{Y}$ to $\mathcal{X}$:

Lemma 2.9. *The operator* (14) *is isometrically isomorphic to the restricted projection operator*

$$P_{\mathcal{Y}|\mathcal{X}} \,:\, L^2(\mathcal{X}\times\mathcal{Y}, \mathcal{X}\times\mathcal{B}(\mathcal{Y}), \mu) \longrightarrow L^2(\mathcal{X}\times\mathcal{Y}, \mathcal{B}(\mathcal{X})\times\mathcal{Y}, \mu), \quad \tilde{f} \longmapsto P\tilde{f}, \tag{25}$$

for the orthogonal projection $P : L^2(\mathcal{X}\times\mathcal{Y}, \mathcal{B}(\mathcal{X}\times\mathcal{Y}), \mu) \twoheadrightarrow L^2(\mathcal{X}\times\mathcal{Y}, \mathcal{B}(\mathcal{X})\times\mathcal{Y}, \mu)$.

Proof. See Appendix B.1.3. (Here, $\mathcal{X}\times\mathcal{B}(\mathcal{Y}) \equiv \{\mathcal{X}\times A \mid A \in \mathcal{B}(\mathcal{Y})\}$, and likewise for $\mathcal{B}(\mathcal{X})\times\mathcal{Y}$.) □

This paper combines the classical variational perspective of Lemma 2.9 with a specific 'feature-based parametrisation' of the form (17) to develop two universally consistent estimators for conditional dependencies between jointly distributed, multidimensional stochastic processes. These estimators, of the type (24), are derived from an ensemble of core rough path theory objects and enjoy a clear mathematical and conceptual interpretation (Section 4). They are also computationally efficient and supported by broad theoretical guarantees, including explicit convergence rates (Section 5).

3 Time-Dependent Data and the Signature Transform

So far, we have not specified the regressor-regressand relation $X \sim Y$ beyond its inclusion in the product $\mathcal{X}\times\mathcal{Y}$ for general Polish spaces $\mathcal{X}$ and $\mathcal{Y}$. In this section, we sharpen this assumption and focus on learning statistical relationships between multidimensional time-dependent data. Specifically, we consider conditional dependencies (8) pertaining to (sufficiently regular) subspaces

$$\mathcal{X} \text{ of } C([0,1];\mathbb{R}^{d_1}) \quad \text{and} \quad \mathcal{Y} \text{ of } C([0,1];\mathbb{R}^{d_2}), \qquad \text{for any } d_1, d_2 \in \mathbb{N}. \tag{26}$$

This setting will allow for an effective and rigorous description of time-dependent statistical relationships betweens multidimensional covariates (X, Y) that are genuinely probabilistic, or modelled as such due to their inherent complexity, and exhibit nuanced temporal and spatial dynamics.

Suitably large data spaces $\mathcal{X}$ and $\mathcal{Y}$ as in (26) are obtained in Section 3.1 as $\|\cdot\|_\infty$-dense linear Polish subspaces of continuous paths by imposing some natural continuity assumptions on the points in $C([0,1];\mathbb{R}^d)$. These data spaces $\mathcal{X}$ and $\mathcal{Y}$, when adequately normed, are effectively (topological) 'Hilbert manifolds' as they admit single, globally-defined coordinate charts:

$$q_{\mathcal{X}} : \mathcal{X} \longrightarrow \mathcal{H}_{\mathcal{X}} \qquad \text{and} \qquad q_{\mathcal{Y}} : \mathcal{Y} \longrightarrow \mathcal{H}_{\mathcal{Y}} \tag{27}$$

mapping the original data spaces $\mathcal{X}$ and $\mathcal{Y}$ into structurally well-behaved Hilbert spaces $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$. In the language of RKHS, the functions (27), called '(bounded) signature transforms' and defined in Sections 3.2 and 3.3, can be viewed as highly structured feature maps for multidimensional sequential data. Serving the role of (15), these maps integrate seamlessly into the context of (10), allowing for a fruitful extension of the general approximation approach (17) (see Sections 4 and 5).

3.1 Spaces of Sequential Data and Stochastic Processes

We specialise the general setting of Section 2 to the case of time-dependent data, and ensure the measure-theoretic adequacy of the resulting state spaces $\mathcal{X}$ and $\mathcal{Y}$ for immediate integration into the framework of Section 2.3. To this end, we define a time-dependent, or 'sequential', datum $z = (z_t)$ as a continuously ordered family of vectors,

$$z \coloneqq (z_t \mid t \in I) \equiv (z_t)_{t\in I} \qquad \text{with} \qquad z_t \equiv (z_t^1, \dots, z_t^d)^\intercal \in \mathbb{R}^d, \tag{28}$$

that is, a continuous map $z : I \to \mathbb{R}^d$ over some fixed compact interval $I \subset \mathbb{R}$, usually $I = [0,1]$. The probabilistic description of time-dependent information (28) is largely a 'macroscopic' endeavour, and so we view the objects (28) primarily as (highly-structured) points of the so-called path spaces

$$C(I;Z) \coloneqq \big\{z \equiv (z_t) \in Z^I \mid I \ni t \mapsto z_t \text{ is continuous}\big\}, \quad \text{normed by} \quad \|z\|_\infty \coloneqq \sup_{t\in I}|z_t|, \tag{29}$$

where $(Z, |\cdot|)$ is a given normed space, usually $Z = \mathbb{R}^d$ for some $d \in \mathbb{N}$; we write $\mathcal{C}_d \coloneqq C([0,1];\mathbb{R}^d)$.

Following on from the general spaces (29), choose $I = [0,1]$ wlog and consider the Banach space

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

For simplicity, this paper operates on the 'smooth core' $\mathcal{C}_d^1$ of absolutely continuous paths in $\mathcal{C}_d$,

$$\mathcal{C}_d^1 \coloneqq \left\{ z \in \mathcal{C}_d \;\middle|\; \exists\, \dot{z} \in L^1([0,1];\mathbb{R}^d) :\; z_\cdot = z_0 + \int_0^\cdot \dot{z}_s \,\mathrm{d}s \right\}, \quad \text{norm} \quad \|z\|_{1\text{-var}} \coloneqq |z_0| + \|\dot{z}\|_{L^1}, \tag{30}$$

the so-called 1-*variation norm*. We note that the dense subspace $\mathcal{C}_d^1$ of $\mathcal{C}_d$ is chosen for technical convenience. However, as usual in rough path theory, this choice incurs essentially no loss of generality, as everything that follows can be extended canonically to spaces of rougher paths [18].

Remark 3.1 (Discrete-Time Data)**.** Note that the above setting covers both continuous- and discrete-time data. Discrete-time data can be naturally embedded in the path spaces (30) through order-preserving piecewise-linear interpolation of discretely observed sequential information. For details on this embedding, see e.g. [49, Section B.2]. ◆

3.1.1 Measure-Theoretical Preliminaries for Path Spaces

Next, let us note the following convenient topological and measure-theoretical properties.

Lemma 3.2. *The space $\mathcal{C}_d^1$ is a Borel subset of $(\mathcal{C}_d, \|\cdot\|_\infty)$ and a separable Banach space wrt. the norm $\|\cdot\|_{1\text{-var}}$, 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 shown in Appdx. B.2.2. □

Specifying the sequential setup (26), fix any predictor- and response-dimension $d_\mathcal{X}, d_\mathcal{Y} \in \mathbb{N}$ and set

$$\mathcal{X} \coloneqq \big(\mathcal{C}_{d_\mathcal{X}}^1, \|\cdot\|_{1\text{-var}}\big) \qquad \text{and} \qquad \mathcal{Y} \coloneqq \big(\mathcal{C}_{d_\mathcal{Y}}^1, \|\cdot\|_{1\text{-var}}\big). \tag{31}$$

For a topology on the Cartesian product $\mathcal{X} \times \mathcal{Y}$ hosting all relations between the points in (31), let

$$\|(x,y)\|_{1\text{-var}} \coloneqq |(x_0,y_0)| + \sup_{\mathcal{D}} \sum_{(t_\nu)\in\mathcal{D}} |(x_{t_\nu}, y_{t_\nu}) - (x_{t_{\nu-1}}, y_{t_{\nu-1}})| \quad \text{and} \quad \|(x,y)\|_\infty \coloneqq \sup_{t\in[0,1]} |(x_t, y_t)|,$$

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

Remark 3.3. Lemma 3.4 below asserts that fusing (31) to the product $(\mathcal{X} \times \mathcal{Y}; \|\cdot\|_{1\text{-var}})$ bears no measure-theoretic complications, confirming that the sequential setting (31) integrates smoothly into the framework discussed in Section 2. Alongside Lemma 3.2, this lemma further ensures that the choice of path-space topology—whether $\|\cdot\|_{1\text{-var}}$ or $\|\cdot\|_\infty$—is inconsequential when working with the Borel measure spaces $\mathcal{M}_1(\mathcal{X})$, $\mathcal{M}_1(\mathcal{Y})$, or $\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, since both topologies induce identical σ-algebras on the respective data spaces $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{X} \times \mathcal{Y}$. ◆

Lemma 3.4. *The space $\mathcal{Z} \coloneqq \big(\mathcal{X} \times \mathcal{Y}, \|\cdot\|_{1\text{-var}}\big)$ is Polish, and its Borel-σ-algebra $\mathcal{B}(\mathcal{Z}) \equiv \mathcal{B}(\mathcal{Z}, \|\cdot\|_{1\text{-var}})$ coincides with $\mathcal{B}(\mathcal{Z}, \|\cdot\|_\infty)$. For maps $X : (\Omega, \mathscr{F}) \to \mathcal{X}$ and $Y : (\Omega, \mathscr{F}) \to \mathcal{Y}$ on a measurable space $(\Omega, \mathscr{F})$, the joint process (X, Y) is $\big(\mathscr{F}, \mathcal{B}(\mathcal{Z})\big)$-measurable iff X and Y are Borel-measurable.*

Proof. See Appendix B.2.3. □

As for the lift of a measure $\mu \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ to the law of an $\mathcal{X} \times \mathcal{Y}$-valued pair (X, Y) of random variables, as described in (11), we as usual define a (continuous-time) *stochastic process* as any map

$$S : \Omega \to C(I; Z) \quad \text{such that} \quad \omega \mapsto S(\omega) \equiv (S_t(\omega))_{t\in I} \quad \text{is } \big(\mathscr{F}, \mathcal{B}(C(I;Z))\big)\text{-measurable}, \tag{32}$$

where the above is defined over some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ (usually left implicit).

Some useful technical details on stochastic processes (32) are collected in Remark B.3.

3.1.2 Example Applications: Stochastic Process Prediction and Classification

As a brief preview of potential future applications, note that the standard regression task of forecasting a stochastic process's future from its past fits seamlessly into the general framework (11):

Any process $S = (S_t(\omega)) : I \times \Omega \to Z$ as in (32) generates two types of sub-σ-algebras of $\mathscr{F}$,

$$\sigma(S) \coloneqq \sigma\big(S^{-1}(B) \,\big|\, B \in \mathcal{B}(C(I;Z))\big) \qquad \text{and} \qquad \sigma_{t]}(S) \coloneqq \sigma(S_s \mid s \in I : s \leq t), \tag{33}$$

for any fixed $t \in I$. The following well-known observation reminds us that (33) are closely related. More specifically, Lemma 3.5 ensures that the common regression application of predicting the future evolution of a stochastic process (S) based on its history (i.e., $\sigma(S_s \mid s \leq t_0)$, for some $t_0 \in I$) is an immediate special instance of the general framework (11): Simply choose $(X, Y) \coloneqq (S^{\wedge t_0}, S)$.

Lemma 3.5 (Prediction)**.** *Let $S = (S_t)_{t\in I}$ be as in (32), and $t_0 \in I$. Then we have $\sigma_{t_0]}(S) = \sigma(S^{\wedge t_0})$ for the stochastic process $S^{\wedge t_0} \coloneqq (S_{t\wedge t_0})_{t\in I}$.*

Proof. Directly by definition of $\mathcal{B}(C(I;Z))$ and (33), see e.g. [24, Problem 2.4.4.2 (p. 60)]. □

A second major machine learning task that can be subsumed under the above framework (11) & (10) is the (probabilistic) classification of time-series/stochastic processes: Suppose that any given datum $x \in \mathcal{X}$, modelled as a sample of some abstract process (32) (that is, $x = S(\omega)$ for an $\omega \in \Omega$) is to be assigned a class label $c_x \in \{c_1, \dots, c_k\} \subset \mathbb{R}$. To ensure this assignment is well-defined, let us call the process S *classifiable* (under k-many labels) if there exists a measurable partition $\Omega = \bigsqcup_{i=1}^k E_i$ (with $E_i \in \mathscr{F}$ for all $i \in [k]$, and $E_j \cap E_k = \emptyset$ if $j \neq k$) such that: for each $\omega \in \Omega$ there is $i = i_\omega \in [k]$ with $S^{-1}(S(\omega)) \subseteq E_i$. We then define $C \coloneqq \sum_{i=1}^k c_i \mathbb{1}_{E_i} : \Omega \to \mathbb{R}$ and call this the class variable associated to S. The following lemma ensures that classification can be treated as a special instance[3] of the regression framework (11) & (10) (recalling (13)): Simply choose $(X, Y) \coloneqq (S, C)$.

Lemma 3.6 (Classification)**.** *Let S be classifiable and C its associated class variable. Then the class labelling $\lambda : \mathcal{X} \ni x \mapsto c_x \in \{c_i\}$ is well-defined on $S(\Omega)$, the conditional dependence of $\mu \coloneqq \mathbb{P}_{(S,C)}$ wrt. $\mathbb{P}_S$ is $\mathfrak{c}_S(\mu) = (\delta_{\lambda(x)} \mid x \in S(\Omega))$, and we have $\lambda = \arg\max_{c\in C(\Omega)} \mathbb{P}(C = c \mid S = \cdot)$ ($\mathbb{P}_S$-a.e.).*

Proof. Let there be k-many class labels, say $c_1, \dots, c_k \in \mathbb{R}$. Since S is classifiable, we know that for each $x \in S(\Omega)$ there is a unique $i_x \in [k]$ with $S^{-1}(\{x\}) \subseteq E_{i_x}$. This defines a function $\varphi : S(\Omega) \to [k]$, $x \mapsto i_x$, which in turn determines the class labelling λ to $\lambda : S(\Omega) \ni x \mapsto c_{\tau(\varphi(x))}$, for $\tau : [k] \xrightarrow{\sim} [k]$ some fixed permutation; this proves the first claim. Next, note that, by construction, $C = \lambda(S)$ pointwise on Ω (assuming $\tau = \mathrm{id}$ wlog). This shows $\mathfrak{c}_S(\mu) = (\delta_{\lambda(x)})$, cf. Example 2.1, and also: $\mathbb{P}(C = c \,|\, S) = \mathbb{E}[\mathbb{1}_{\{c\}}(\lambda(S)) \,|\, S] = \mathbb{1}_{\{c\}}(\lambda(S)) = \delta_{c,\lambda(S)}$, which proves the last claim. □

3.2 The Signature Representation of Time-Dependent Data

We proceed to define the (bounded) signature transforms (27) over paths [9, 35, 12], which are coordinate charts that concisely and efficiently embed the previously introduced data domains (31) into (bounded subsets of) well-organized Hilbert spaces. These charts are conveniently computable, continuous, and graded injections, formulated as multivariate formal power series with coefficients given by rapidly decaying iterated integrals. As later sections will show, the rich structure of these charts, carefully discussed below, positions them as valuable tools for a precise stochastic analysis of time-dependent conditional dependencies (9) within a streamlined functional analytic context.

[3] Note that the class variable C lifts to a stochastic process $C : \Omega \to \mathcal{C}_1^1$ under the trivial embedding $\mathbb{R} \hookrightarrow \mathcal{C}_1^1$ which identifies the reals with constant paths.

3.2.1 Definition of the Signature

At the most basic level, the signature transform is a faithful (i.e., one-to-one) compression that maps a path to a hierarchically graded list of countably many numerical coordinates of that path.

A convenient indexing of these coordinates requires some basic 'multiindex notation':

Notation 3.7 (Words and Formal Power Series)**.** Let $d \in \mathbb{N}$ be fixed. We denote by

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

the free monoid over the alphabet $[d] := \{\mathtt{1}, \mathtt{2}, \ldots, \mathtt{d}\}$, representing all finite sequences (or 'words') of zero or more elements from $[d]$ (the 'letters'), with $\emptyset$ symbolizing the empty word. The space $\mathbb{R}[[d]]$ ($\cong \mathbb{R}[[\mathtt{x}_1, \ldots, \mathtt{x}_d]]$) denotes the (free) algebra of all multivariate formal power series in the non-commutative variables $\mathtt{1}\,(\cong \mathtt{x}_1), \ldots, \mathtt{m}\,(\cong \mathtt{x}_d)$, with $\mathbf{1} := 1 \cdot \emptyset$ its multiplicative unit. Explicitly,

$$\mathbb{R}[[d]] = \{\boldsymbol{t} : [d]^* \to \mathbb{R} \mid \boldsymbol{t} \text{ is a map}\} \equiv \left\{ \textstyle\sum_{w \in [d]^*} \boldsymbol{t}_w \cdot w \cong (\boldsymbol{t}_w)_{w \in [d]^*} \;\middle|\; \boldsymbol{t}_w \in \mathbb{R} \right\} \cong \prod_{\nu=0}^{\infty} (\mathbb{R}^d)^{\otimes \nu}, \tag{34}$$

where each word $\mathtt{i}_1 \cdots \mathtt{i}_d \in [d]^*$ is identified with its associated elementary tensor $e_1 \otimes \cdots \otimes e_d \in (\mathbb{R}^d)^{\otimes m}$, writing $(e_i)_{i \in [d]}$ for the standard basis in $\mathbb{R}^d$. The length $|w|$ of a word w is defined as the number of its letters, so that any length-k word ($k \in \mathbb{N}$) is of the form

$$w = \mathtt{i}_1 \mathtt{i}_2 \cdots \mathtt{i}_{k-2} \mathtt{i}_{k-1} \mathtt{i}_k \in [d]^*, \quad \text{with} \quad \mathtt{i}_\nu \in \{\mathtt{1}, \ldots, \mathtt{d}\} \;\text{ for each } \nu \in [k]. \tag{35}$$

All of this extends to the alphabet $[\underline{d}]^*$ with $[\underline{d}] := \{0, 1, \ldots, d\}$. For any $k \in \mathbb{N}$, we write $\Delta_k := \{(t_\nu) \in [0,1]^k \mid 0 \le t_1 \le t_2 \le \cdots \le t_k \le 1\}$ for the k-dimensional standard simplex. ◆

Take any path $z \equiv (z^1, \cdots, z^d) \in \mathcal{C}_d^1$ with components $z^i \in \mathcal{C}_1^1$. For compact notation, let us for any word $w \equiv \mathtt{i}_1 \mathtt{i}_2 \cdots \mathtt{i}_k \in [d]^*$ ($k \in \mathbb{N}$) consider the w-indexed differential k-form

$$\mathrm{d}z^w := \mathrm{d}z^{i_1} \wedge \mathrm{d}z^{i_2} \wedge \cdots \wedge \mathrm{d}z^{i_k} = \dot{z}_{t_1}^{i_1} \dot{z}_{t_2}^{i_2} \cdots \dot{z}_{t_k}^{i_k} \, \mathrm{d}t_1 \wedge \mathrm{d}t_2 \wedge \cdots \wedge \mathrm{d}t_k.$$

The following map is (essentially) a global coordinate chart that elucidates the spaces (31) of sequential data by embedding their elements into a Hilbert space where they are easier to analyse.

Definition 3.8 (Signature)**.** The *signature* $\mathfrak{sig} : \mathcal{C}_d^1 \to \mathbb{R}[[d]]$ sends a path z to the formal power series

$$\mathfrak{sig}(z) := \sum_{w \in [d]^*} \int_{\Delta_{|w|}} \mathrm{d}z^w \cdot w \cong \left(\int_{\Delta_{|w|}} \mathrm{d}z^w \;\middle|\; w \in [\mathtt{1}, \ldots, \mathtt{d}]^* \right). \tag{36}$$

Written out, the w-th signature coefficient of a path $z = (z_t^1, \cdots, z_t^d)_{t \in [0,1]}$, denoted $\mathfrak{sig}_w(z)$, reads

$$\int_{\Delta_{|w|}} \mathrm{d}z^w \equiv \int_0^1 \int_0^{t_k} \int_0^{t_{k-1}} \cdots \int_0^{t_3} \int_0^{t_2} \mathrm{d}z_{t_1}^{i_1} \mathrm{d}z_{t_2}^{i_2} \cdots \mathrm{d}z_{t_{k-2}}^{i_{k-2}} \mathrm{d}z_{t_{k-1}}^{i_{k-1}} \mathrm{d}z_{t_k}^{i_k}, \tag{37}$$

for any word $w \in [d]^*$ of the general form (35); these are each iterated Lebesgue-Stieltjes integrals.

3.2.2 Some Basic Properties of the Signature

Transitioning from a path's trace to its graph, let us now map a path $z \equiv (z^1, \cdots, z^d) \in \mathcal{C}_d$ to

$$\bar{z} := (t, z_t)_{t \in [0,1]} \equiv \left(z_t^0, z_t^1, \cdots, z_t^d\right)_{t \in [0,1]} \in \mathcal{C}_{d+1}. \tag{38}$$

The mapping $\bar{\iota} : z \mapsto \bar{z}$ embeds $\mathcal{C}_d^1$ into $\mathcal{C}_{d+1}^1$, and composing it with (36) results in an actual embedding of paths into the space of formal power series.

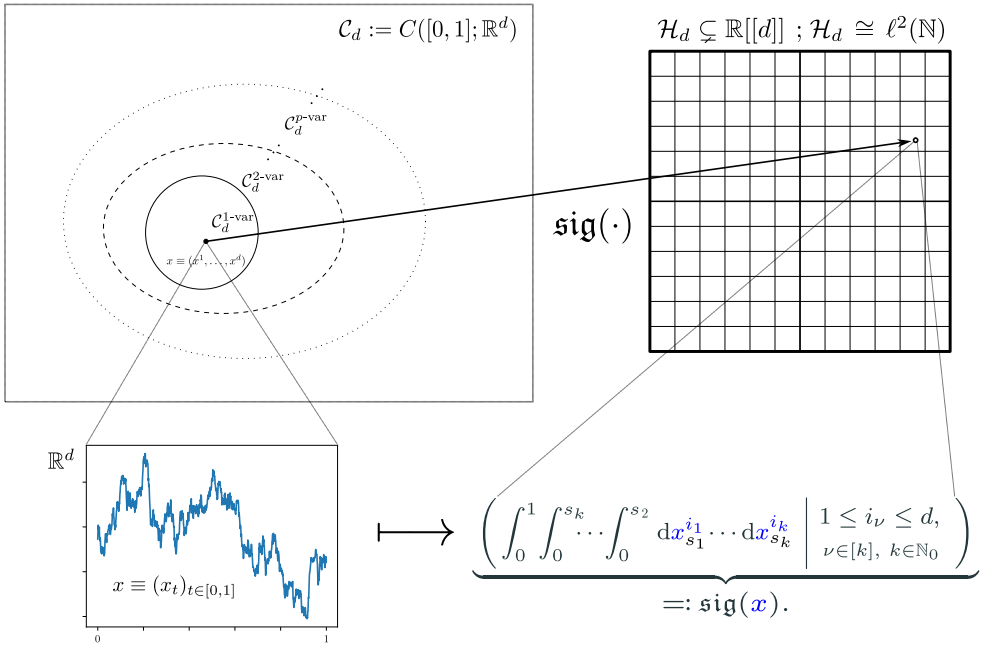

Figure 2: *The signature, $\mathfrak{sig}$, is a global Hilbert-valued coordinate chart on (sufficiently continuous) time-dependent data that maps a multidimensional path to an ordered list of its iterated integrals.*

Theorem 3.9 ([20])**.** *The augmented signature map*

$$\underline{\mathfrak{sig}} := \mathfrak{sig} \circ \bar{\iota} \;:\; \mathcal{C}_d^1 \longrightarrow \mathbb{R}[[\underline{d}]], \quad z \mapsto \mathfrak{sig}(\bar{z}), \quad \textit{is injective.} \tag{39}$$

Proof. Immediate by [20, 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 $x = y$. □

To exploit the transform (36), or rather its augmented version (39), for our analysis, we need to enhance its co-domain with additional analytic structure. Specifically, we see next that a subspace of $\mathbb{R}[[d]]$ [resp. $\mathbb{R}[[\underline{d}]]$] which includes the range $\mathrm{im}(\mathfrak{sig})$ [resp. $\mathrm{im}(\underline{\mathfrak{sig}})$] of the [augmented] signature, can be easily structured as a Hilbert space. This requires some preliminary basic notation:

Notation 3.10 (Gradation, Projections, and Inner Products)**.** Any $[d]^*$-indexed infinite tuple $\boldsymbol{a} \equiv (a_w)_{w\in[d]^*} \subset \mathbb{R}$, such as (36), can be injected into the set (34) via $\boldsymbol{a} = \sum_{w\in[d]^*} \boldsymbol{t}_a(w)\cdot w \in \mathbb{R}[[d]]$, where $\boldsymbol{t}_a(\mathtt{i}_1 \cdots \mathtt{i}_m) := a_{i_1\cdots i_m}$. Upon grouping the summands in (34) by their wordlength, we get

$$V \equiv \mathbb{R}[[d]] = \prod_{m=0}^{\infty} V_m \qquad \text{with} \qquad V_m := \bigoplus_{w\in[d]^* \,:\, |w|=m} \mathbb{R}w \tag{40}$$

for $V_0 := \mathbb{R}$. The decomposition (40) of $\mathbb{R}[[d]]$ into an infinite product of homogeneous components (V_m) defines a *gradation* of V, accompanied by projections $\pi_m : V \to V_m$, $\boldsymbol{a} \mapsto \pi_m(\boldsymbol{a}) := \boldsymbol{a}_m \equiv \sum_{|w|=m} a_w \cdot w$; we set $\pi_{[m]} \equiv \sum_{\nu=1}^m \pi_\nu \;:\; V \longrightarrow V_{[m]} := \bigoplus_{j=0}^m V_j \subset V$. Finally, the *inner product*

$$\langle \cdot\,, \cdot \rangle \;:\; V \times V \;\to\; \overline{\mathbb{R}}, \quad (\boldsymbol{a}, \boldsymbol{b}) \mapsto \sum\nolimits_{w\in[\boldsymbol{d}]^*} \langle \boldsymbol{a}, w\rangle \cdot \langle \boldsymbol{b}, w\rangle =: \langle \boldsymbol{a}, \boldsymbol{b}\rangle, \tag{41}$$

is defined as the (infinite) bilinear extension of $\langle u, v\rangle := \delta_{uv}$, $u, v \in [d]^*$. In other words, $[d]^*$ 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 $V \times V$. ◆

Example 3.11. To illustrate the above notation, note for instance that, for $d \geq 4$ say,

$$3\,\mathfrak{sig}_{21}(x) - 25\,\mathfrak{sig}_{44}(x) = \langle 3\cdot 21 - 25\cdot 44, \mathfrak{sig}(x)\rangle = 3\int_0^1\int_0^t \mathrm{d}x^2_s\mathrm{d}x^1_t - 25\int_0^1\int_0^t \mathrm{d}x^4_s\mathrm{d}x^4_t,$$

$$\langle 10230, \underline{\mathfrak{sig}}(x)\rangle = \int_{\Delta_5} \mathrm{d}x^1 \wedge \mathrm{d}s \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3 \wedge \mathrm{d}t = \int_0^1\int_0^{t_5}\int_0^{t_4}\int_0^{t_3}\int_0^{t_2} \dot{x}^1_{t_1} t_2 \dot{x}^2_{t_3}\dot{x}^3_{t_4} t_5 \,\mathrm{d}t_1\mathrm{d}t_2\mathrm{d}t_3\mathrm{d}t_4\mathrm{d}t_5,$$

and further $\pi_4(1232 + 2\cdot 243 - 0.5\cdot 22 - 3\cdot 2334) = 1232 - 3\cdot 2334$, $\pi_2(231) = 0\,(\equiv 0\cdot\emptyset)$, $\pi_{[3]}(1 - 32 + 1345 + 6\cdot 333 + 2\cdot 32) = 1 + 32 + 6\cdot 333$, and $\langle 3\cdot 11 - 2\cdot 132, 123 - 4\cdot 11\rangle = -12$. ◆

The structure (41) allows us to configure V as a Hilbert space, under convergence constraints:

Lemma 3.12 (Hilbert Codomain of (36))**.** *For $(V, \langle\cdot,\cdot\rangle)$ the power series algebra from above, let*

$$\mathcal{H}_d := \left\{ \boldsymbol{t} \in V \;\middle|\; \|\boldsymbol{t}\| := \sqrt{\textstyle\sum_{m\geq 0}\|\pi_m(\boldsymbol{t})\|^2_m} < \infty \right\} \quad \textit{with} \quad \|\cdot\|_m := \sqrt{\langle\cdot,\cdot\rangle_m}\,, \tag{42}$$

where $\langle\cdot,\cdot\rangle_m := \langle\pi_m(\cdot), \pi_m(\cdot)\rangle$ for each $m \geq 0$. Then $(\mathcal{H}_d, \langle\cdot,\cdot\rangle)$ is a separable Hilbert space with orthonormal basis $(w \mid w \in [\boldsymbol{d}]^*)$, containing the image $\mathfrak{sig}(\mathcal{C}^1_d)$ and all of $\mathbb{R}[d]$. Moreover, the maps

$$\mathfrak{sig} : \mathcal{C}^1_d \to \mathcal{H}_d \qquad \textit{and} \qquad \underline{\mathfrak{sig}} : \mathcal{C}^1_d \to \mathcal{H}_{\underline{d}} \tag{43}$$

are both continuous wrt. the p-variation topology, for any $p \geq 1$.

The statements of this lemma are all well-known, but see Appendix B.2.4.

Remark 3.13. By weighting the graded components (V_m) in (40) before ℓ^2-direct-summing them into a composite Hilbert space as in (42), we can generalise the standard space $\mathcal{H}_d$ to a whole family of alternate $\mathfrak{sig}$- resp. $\underline{\mathfrak{sig}}$-containing Hilbert codomains $\mathcal{H}^\gamma_d$. See Remark B.4 for details. ◆

Geometrically, the (augmented) signature map $\underline{\mathfrak{sig}}$ from (43) serves as a global coordinate chart[4] for the Hilbert manifold $\mathcal{C}^1_d$. Similar to general coordinate charts that map from a less intuitive space (like (30)) into a more comprehensible 'analysis space' (such as (42)), we will use the signature representation (36) of the paired spaces (31) to facilitate the analysis of statistical dependencies between sequential data through (17) and (27) via the lens (43).

To this end, the next subsection bridges the gap between (43) and the desired transforms (27).

3.3 Bounded Signature Transforms

The primary aim of this paper is to propose a universally consistent estimator for conditional dependencies in sequential data, as detailed in Definition 2.6. Employing the composite approach (24), as suggested by (17), requires the square integrability, with respect to any measure in $\mathcal{M}_1(\mathcal{Y})$, of the (31)-tailored coordinate chart $q_{\mathcal{Y}}$ from (27) (whose prototype is (15)). The only way to ensure this is for the function $q_{\mathcal{Y}}$ to be bounded. This is achieved by the concepts of tensor normalization and 'robust signatures' introduced in [12], which this section implements to confine the signature transforms (43) to a bounded subdomain of (42) in a structure-preserving fashion. This finalizes the transition from the default charts (43) to the (15)-respecting enhancements (27).

To begin, let us state some basic facts on the growth [wrt. their gradation index m] and continuity of the coordinate representations (43). This section consistently applies Notation 3.10.

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, $\delta_1 = \mathrm{id}_V$).

[4] Notwithstanding the continuity of its inverse $\underline{\mathfrak{sig}}^{-1}$, which does not concern us in this paper.

Proposition 3.14 (Signature Decay; e.g. [36, Theorem 3.7 (case $p = 1$)])**.** *We have that*

$$\big\|\pi_m(\mathfrak{sig}(x))\big\|_m \;\le\; \|x\|_{1\text{-var}}^m/(m!\,\beta), \quad \textit{for each} \;\; (x,m) \in \mathcal{C}_d^1 \times \mathbb{N}_0, \tag{44}$$

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

$$\mathfrak{sig}(\mathcal{C}_d^1) \;\subset\; \mathcal{H}_d^{\downarrow} \coloneqq \Big\{ \boldsymbol{t} \in V \;\Big|\; \textstyle\sum_{m\ge 0} \|\pi_m(\boldsymbol{t})\|_m \lambda^m < \infty,\; \forall\, \lambda > 0 \Big\} \;\subseteq\; \mathcal{H}_d,$$

and consequently $\delta_\lambda(\mathfrak{sig}(\mathcal{C}_d^1)) \subset \mathcal{H}_d^{\downarrow}$ for each $\lambda > 0$, since clearly $\delta_\lambda : \mathcal{H}_d^{\downarrow} \to \mathcal{H}_d^{\downarrow}$ for each $\lambda > 0$.

Lemma 3.15 ('Strong Continuity')**.** *Endow the subspace $\mathcal{H}_d^{\downarrow}$ with the locally convex topology $\tau_{\downarrow}$ that is induced by the family of norms*

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

Then $(\mathcal{H}_d^{\downarrow}, \tau_{\downarrow})$ is separable and metrizable Hausdorff and, for each $p \ge 1$, the signature transform

$$\mathfrak{sig} \;:\; \big(\mathcal{C}_d^1, \|\cdot\|_{p\text{-var}}\big) \;\to\; (\mathcal{H}_d^{\downarrow}, \tau_{\downarrow}) \quad \textit{is continuous.} \tag{45}$$

Proof. The topological qualities of $(\mathcal{H}_d^{\downarrow}, \tau_{\downarrow})$ are due to [11, Corollary 2.4], while the continuity assertion about the signature is contained in [11, Corollary 5.5]. □

Remark 3.16 (Comparison of Topologies)**.** Clearly, for any topological space $\mathcal{T}$, a map $\varphi : \mathcal{H}_d^{\downarrow} \to \mathcal{T}$ is $\tau_{\downarrow}$-continuous if φ is $|||\cdot|||_\lambda$-continuous for at least one $\lambda > 0$. Moreover, $\mathcal{H}_d^{\downarrow}$ is a subspace of each $\mathcal{H}_d^{\gamma}$ (Remark B.4), and the above topology $\tau_{\downarrow}$ on $\mathcal{H}_d^{\downarrow}$ is *finer* than the (127)-induced subspace topology τ_γ on $\mathcal{H}_d^{\downarrow}$ (Appendix B.2.5). Thus, statement (45) also holds for $\mathcal{H}_d^{\downarrow}$ replaced by $\mathcal{H}_d^{\gamma}$. ◆

Lemma 3.17. *If $\lambda_\cdot : (\mathcal{H}_d^{\downarrow}, \tau_{\downarrow}) \to \mathbb{R}_{>0}$ is a continuous positive scalar field, then the map*

$$\Lambda \;:\; (\mathcal{H}_d^{\downarrow}, \tau_{\downarrow}) \to (\mathcal{H}_d^{\downarrow}, \|\cdot\|), \quad \boldsymbol{t} \mapsto \delta_{\lambda_{\boldsymbol{t}}} \boldsymbol{t}, \quad \textit{is continuous.} \tag{46}$$

Consequently and for any fixed $p \ge 1$, the Λ-scaled augmented signature transform

$$\Lambda \circ \underline{\mathfrak{sig}} \;:\; (\mathcal{C}_d^1, \|\cdot\|_{p\text{-var}}) \;\to\; (\mathcal{H}_{\underline{d}}^{\downarrow}, \|\cdot\|), \quad x \;\mapsto\; \delta_{\lambda_{\underline{\mathfrak{sig}}(x)}}(\underline{\mathfrak{sig}}(x)), \quad \textit{is continuous.} \tag{47}$$

Proof. See Appendix B.2.6 □

With the above observations in mind, we can now proceed to modify the Hilbert charts (43) to ensure they are integrable with respect to any (Borel) probability measure on their respective domains (31), which serve as our regression spaces. The idea, originally proposed in [12], is to continuously inject ('squish') the image $\underline{\mathfrak{sig}}(\mathcal{C}_d^1)$ of the chart (39) into a ball of finite radius in $\mathcal{H}_{\underline{d}}$ by composing the chart $\underline{\mathfrak{sig}}$ with a bounded 'squeezing dilation' of the form (46). Provided that the chosen dilation (46) is also a continuous injection, this ensures that the squished signature (47) is a Hilbert chart of the envisioned kind (27). The next definition summarizes this procedure.

Definition 3.18 (cf. [12])**.** We call *feature normalisation* (fN) any injective map of the form

$$\Lambda \;:\; \mathcal{H}_{\underline{d}}^{\downarrow} \to \mathcal{H}_{\underline{d}}^{R} \coloneqq \{\boldsymbol{t} \in \mathcal{H}_{\underline{d}} \mid \|\boldsymbol{t}\| \le R\}, \quad \boldsymbol{t} \mapsto \delta_{\lambda_{\boldsymbol{t}}}\, \boldsymbol{t}, \tag{48}$$

for $R > 0$ some fixed constant and $\lambda : (\mathcal{H}_{\underline{d}}^{\downarrow}, \tau_{\downarrow}) \ni \boldsymbol{t} \mapsto \lambda_{\boldsymbol{t}} \in \mathbb{R}_{>0}$ continuous. Given an fN Λ, we call

$$\underline{\mathfrak{sig}}_\Lambda \coloneqq \Lambda \circ \underline{\mathfrak{sig}} \;:\; \mathcal{C}_d^1 \longrightarrow \mathcal{H}_{\underline{d}}^{R_\Lambda} \tag{49}$$

a (Λ-)*bounded signature transform* (here, $R_\Lambda \coloneqq \sup\{\|\Lambda(\boldsymbol{t})\| \,|\, \boldsymbol{t} \in \mathcal{H}_{\underline{d}}^{\downarrow}\}$).

[12, Section 3.2] show that feature normalisations exist and can be conveniently constructed. Furthermore, [12] show that the specific form (46) of these transforms preserves certain algebraic properties of $\underline{\mathfrak{sig}}$, which is crucial for proving that bounded signatures fulfil the universality property (15); see Section 4 below. The following result is a slight extension of [12, Theorem 21].

Proposition 3.19 (cf. [12, Theorem 21])**.** *Let $\Lambda = \delta_{\lambda_{(\cdot)}}$ be an fN with scalar field $\lambda_\cdot : (\mathcal{H}^{\downarrow}_{\underline{d}}, \tau_{\downarrow}) \to \mathbb{R}_{>0}$, and set $\underline{\lambda}_{(\cdot)} \coloneqq \lambda_\cdot \circ \underline{\mathfrak{sig}}(\cdot)$. Further, abbreviate $\mathcal{Z} \coloneqq (\mathcal{C}^1_d, \|\cdot\|_{1\text{-var}})$ and denote by*

$$\mathcal{A}_\Lambda \coloneqq \operatorname{span}_{\mathbb{R}}\Big\{ \underline{\xi}^\Lambda_w \,:\, \mathcal{Z} \ni x \mapsto \underline{\lambda}^{|w|}_x \int_{\Delta_{|w|}} \mathrm{d}\bar{x}^w \;\Big|\; w \in [\underline{d}]^* \Big\}, \quad \textit{thus} \quad \underline{\mathfrak{sig}}_\Lambda = \sum\nolimits_{w\in[\underline{d}]^*} \underline{\xi}^\Lambda_w \cdot w, \tag{50}$$

the linear span of the component functions $(\underline{\xi}^\Lambda_w \mid w \in [\underline{d}]^*)$ of the Λ-bounded signature $\underline{\mathfrak{sig}}_\Lambda$. Then $\mathcal{A}_\Lambda$ is a point-separating and non-vanishing[5] subalgebra of $C_b(\mathcal{Z})$, the algebra of all bounded continuous functions on $\mathcal{Z}$. Moreover, $\mathcal{A}_\Lambda$ is dense in $\big(C_b(\mathcal{Z}), \tau^{\mathcal{Z}}_{\mathrm{str}}\big)$, for $\tau^{\mathcal{Z}}_{\mathrm{str}}$ the strict topology[6] on $C_b(\mathcal{Z})$. If instead the domain $\mathcal{Z}$ is a bounded subset of $(\mathcal{C}^1_d, \|\cdot\|_{1\text{-var}})$, then all of the above holds for $\Lambda = \mathrm{id}_{\mathcal{H}^{\downarrow}_{\underline{d}}}$.

Proof. The fact that $\mathcal{A}_\Lambda$ is an algebra is due to the span of all iterated integrals (37) being closed under multiplication, a property preserved by the dilation structure (46). The inclusion $\mathcal{A}_\Lambda \subset C_b(\mathcal{Z})$ follows clearly from (48) and Lemma 3.17. That $\mathcal{A}_\Lambda$ is point-separating is ensured by Theorem 3.9 and the injectivity of Λ, while $\mathcal{A}_\Lambda$ being non-vanishing is evident since $\underline{\xi}^\Lambda_\emptyset = 1$ by definition of (49).

Given the preceding discussion, the statements regarding the strict topology follow immediately from [19, Thm. 3.1], cf. [12, Thm. 21]. All details are provided in Appdx. B.2.7 for completeness. □

Proposition 3.19, i.e. the asserted denseness of $\mathcal{A}_\Lambda$ in $C_b(\mathcal{Z})$ with respect to the strict topology, generalises the Stone-Weierstrass theorem to the non locally-compact setting of path spaces (31). This aspect of (49) holds secondary significance for the main objective of this article and will be utilized merely to substantiate a few more peripheral corollaries in Section 4.3.

4 Signature Representation of Time-Dependent Association

The signature transform, introduced as a global Hilbert-valued coordinate map on spaces of sufficiently continuous paths (Section 3.2), is a powerful and well-structured tool for analysing time-dependent data, and it retains this property when injected into bounded subsets (Section 3.3). With this data-analytic concept in hand, we now return to our main task (10) of estimating statistical dependencies, represented as regression operators (Proposition 2.4), between sets (26) of time-dependent multidimensional data. Our strategy for achieving this is as follows:

The statistical approximation of conditional dependencies (8) relies on the σ-algebra generated by the regressor, which serves as an information reservoir for the approximants of (14). Section 4.1 uses the bounded signature transform to discretise this reservoir (Lemma 4.3) and exhaust its operationalisable information with a system of simple signature-based functions (Proposition 4.5 and Corollary 4.6). These results provide us with a tailor-made architecture (58) for the convex approximation of statistical dependencies, with the gradation index m in (40) of the signature serving as the central control parameter for the resolution of this approximation. This architecture is put to use in Section 4.2, where the conditional expectation of the bounded signature of a process Y given another process X, as well as the conditional expectation of a random vector Z given X, are characterised as the solutions of conveniently approximable, convex semi-infinite linear least squares problems (Theorem 4.8 and Corollary 4.9). These solutions yield 'variational signature representations' of the respective conditional expectations, which are then extended in Section 4.3 to asymptotic variational representations (17) of the regression operator (14) (Proposition 4.10).

[5] A family $\mathcal{A} \subseteq C(\mathcal{Z})$ will be called *non-vanishing* if: $\forall\, x \in \mathcal{Z}$ there exists $\varphi \in \mathcal{A}$ such that $\varphi(x) \neq 0$. [6] See e.g. [19] or, for internal reference, Definition B.5.

In combination, this provides a convex nonparametric regression framework for stochastic processes that makes essentially no assumptions about the joint distribution of regressor and regressand, and allows for a comprehensive statistical estimation theory and error control (Section 5).

4.1 A Dense Discretisation of Time-Dependent Information

To begin, we re-instantiate the sequential setting of Section 3.1, meaning that from here onwards:

$$\mathcal{X} \coloneqq \big(\mathcal{C}^1_{d_\mathcal{X}}, \|\cdot\|_{1\text{-var}}\big) \quad \text{and} \quad \mathcal{Y} \coloneqq \big(\mathcal{C}^1_{d_\mathcal{Y}}, \|\cdot\|_{1\text{-var}}\big) \qquad \text{for some} \quad d_\mathcal{X}, d_\mathcal{Y} \in \mathbb{N}, \tag{51}$$

denoting $(\underline{d}, \tilde{\underline{d}}) \coloneqq \big(\underline{d_\mathcal{X}}, \underline{d_\mathcal{Y}}\big)$ for brevity. Then, any probability measure $\mu \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ admits a lift $\mu = \mathbb{P}_{(X,Y)}$ to the law of some pair of jointly distributed stochastic processes

$$X : \Omega \to \mathcal{X} \quad \text{and} \quad Y : \Omega \to \mathcal{Y} \tag{52}$$

over some (complete) probability space $(\Omega, \mathscr{F}, \mathbb{P})$, see (11) and Remark 2.3. These X and Y induce

$$\text{sub-}\sigma\text{-algebras} \quad \Sigma_X \coloneqq \sigma(X) \quad \text{and} \quad \Sigma_Y \coloneqq \sigma(Y) \quad \text{of} \quad \mathscr{F}.$$

Remark 4.1. Let us include a few basic remarks on the above σ-algebras and measurability.

(i) We recall the set system Σ_X to serve as an information base for the best-approximation (in the Bochner-L^2-distance) $Y_X =: \mathbb{E}[Y \,|\, X]$ of the process Y from within all $\mathcal{Y}$-valued measurable functions of X, cf. also Lemma 2.9. Recall that this best-approximation $\mathbb{E}[Y \,|\, X]$ 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. [48, Theorem II.2.1]. For the 'canonical' probability space $(\Omega, \mathscr{F}, \mathbb{P}) \coloneqq (\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X} \times \mathcal{Y}), \mu)$ (cf. Remark B.1), the sub-σ-algebras Σ_X and Σ_Y of $\mathscr{F}$ correspond to the sub-σ-algebras $\mathcal{B}(\mathcal{X}) \times \mathcal{Y}$ and $\mathcal{X} \times \mathcal{B}(\mathcal{Y})$ of $\mathcal{B}(\mathcal{X} \times \mathcal{Y})$, respectively.

(ii) Throughout, any random variable is understood to be Borel-measurable. In particular: for any Hilbert space $(\mathcal{H}, \|\cdot\|_\mathcal{H})$ and function $\mathbb{Z} : \Omega \to \mathcal{H}$, we call $\mathbb{Z}$ an $\mathcal{H}$-valued random variable ($\Leftrightarrow$: $\mathbb{Z} \in \mathcal{L}^0(\mathbb{P}; \mathcal{H})$) iff $\mathbb{Z}$ is $\big(\mathscr{F}, \sigma(\|\cdot\|_\mathcal{H})\big)$-measurable. The latter measurability is equivalent to $\mathbb{Z}$ being Bochner-measurable thanks to [58, Prop. 1.8], provided that $\mathcal{H}$ is separable. ◆

A best-approximation Y_X of the full-process Y given X is generally difficult to come by, especially due to the 'analytical intractability' of the path space $\mathcal{Y}$. Luckily, such an approximation Y_X is also not directly required: Instead of their Banach-valued default representation (52), it suffices for (10) to examine the covariates X and Y through the lens of their bounded Hilbert-coordinates (49).

So for any two feature normalisations $\Xi : \mathcal{H}^\downarrow_\mathcal{X} \to \mathcal{H}_\mathcal{X}$ and $\Lambda : \mathcal{H}^\downarrow_\mathcal{Y} \to \mathcal{H}_\mathcal{Y}$, see (48), we denote by

$$\mathbb{X}^\Xi \coloneqq \underline{\mathfrak{sig}}_\Xi(X) \qquad \text{and} \qquad \mathbb{Y}^\Lambda \coloneqq \underline{\mathfrak{sig}}_\Lambda(Y) \tag{53}$$

the respective Hilbert-valued representations of X and Y. We then know from (49) that $\mathbb{X}_\Xi$ and $\mathbb{Y}_\Lambda$ are bounded Hilbert-valued random variables, specifically:

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

with $\mathcal{H}_\mathcal{X} \coloneqq \mathcal{H}_{\underline{d}}$ and $\mathcal{H}_\mathcal{Y} \coloneqq \mathcal{H}_{\tilde{\underline{d}}}$ the Hilbert spaces of square-summable power series from (42). Differentiating degrees of integrability $1 \leq p < \infty$, we also introduce the (Bochner-)L^p-spaces

$$L^p(\mathbb{P}; \mathcal{H}) \coloneqq \big\{\mathbb{Z} \in L^0(\mathbb{P}; \mathcal{H}) \;\big|\; \|\mathbb{Z}\|_{L^p(\mathcal{H})} < \infty\big\} \quad \text{for} \quad \big\|\mathbb{Z}\big\|_{L^p(\mathcal{H})} \coloneqq \bigg(\int_\Omega \big\|\mathbb{Z}(\omega)\big\|^p_\mathcal{H} \mathrm{d}\mathbb{P}\bigg)^{1/p} = \mathbb{E}\big[\|\mathbb{Z}\|^p_\mathcal{H}\big]^{\frac{1}{p}}. \tag{54}$$

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

$$\langle \mathbb{Z}_1, \mathbb{Z}_2 \rangle_{L^2(\mathcal{H})} \coloneqq \int_\Omega \langle \mathbb{Z}_1(\omega), \mathbb{Z}_2(\omega) \rangle_{\mathcal{H}} \, \mathrm{d}\mathbb{P} = \mathbb{E}\big[\langle \mathbb{Z}_1, \mathbb{Z}_2 \rangle_{\mathcal{H}}\big]. \tag{55}$$

The boundedness of (53) ensures that there will be no integrability concerns for us here.

Lemma 4.2. *We have that $\mathbb{X}^\Xi \in L^p(\mathbb{P};\mathcal{H}_\mathcal{X})$ and $\mathbb{Y}^\Lambda \in L^p(\mathbb{P};\mathcal{H}_\mathcal{Y})$, for each $p \geq 1$.*

Proof. The inclusion $\mathbb{Y}^\Lambda \in \mathcal{L}^0(\mathbb{P};\mathcal{H}_\mathcal{Y})$ is due to (47) and the fact that in our setting, Bochner- and Borel-measurability coincide thanks to the separability of $\mathcal{H}_\mathcal{Y}$, see Remark 4.1 (ii). The unrestricted integrability of $\mathbb{Y}^\Lambda$ is clear since Λ is bounded by definition (48) of a feature normalisation. □

Let us also make the following simple but useful observation, proved in Appendix B.2.8.

Lemma 4.3. *For each feature normalisation $\Xi : \mathcal{H}_\mathcal{X}^\downarrow \to \mathcal{H}_\mathcal{X}$, we have that*

$$\Sigma_X = \sigma\big(\underline{\mathfrak{sig}}_\Xi(X)\big) = \sigma\Big(\underline{\xi}^\Xi_w(X) \;\Big|\; w \in [\underline{d}]^*\Big) \tag{56}$$

(in the notation of Proposition 3.19).

Let us further introduce the spaces (which for $p = 2$ can be identified with the domain of (25))

$$L^p_X(\mathcal{H}_\mathcal{Y}) \coloneqq L^p(\mathbb{P}, \Sigma_X; \mathcal{H}_\mathcal{Y}) \equiv \{\mathbb{Z} \in L^p(\mathbb{P};\mathcal{H}_\mathcal{Y}) \mid \mathbb{Z}^{-1}(\mathcal{B}(\mathcal{H}_\mathcal{Y})) \subseteq \Sigma_X\} \quad \big(p \in [1,\infty)\big) \tag{57}$$

of all Σ_X-measurable mean-p-integrable $\mathcal{H}_\mathcal{Y}$-valued random variables, and the parameter space

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

The following proposition provides an architecture of readily adjustable model functions through which we can compute the conditional expectation $\mathbb{E}[\mathbb{Y}^\Lambda \,|\, X]$ as the unique solution to a convex and practically implementable 'least-squares type' optimization problem (Theorem 4.8). The proof of this proposition is based on the classical fact that if a class of bounded functions contains all indicator functions of a π-system and is closed under taking pointwise limits of uniformly bounded, monotone sequences, then this class contains all bounded, measurable functions with respect to the σ-algebra generated by the π-system. This fact is formulated as Lemma 4.4 below.

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. See, for instance, [22, Theorem A.1]. □

Proposition 4.5. *For any given feature normalisation $\Xi : \mathcal{H}_\mathcal{X}^\downarrow \to \mathcal{H}_\mathcal{X}$, consider the family*

$$\Psi_X \coloneqq \Big\{\psi_\alpha : \mathcal{X} \to \mathcal{H}_Y \cup \{\infty\} \;\Big|\; \alpha \equiv (\alpha_w) \in \mathfrak{L}^2_{X_\Xi}\Big\}, \quad \textit{with} \quad \psi_\alpha \coloneqq \sum\nolimits_{w\in[\underline{d}]^*} \big\langle \alpha_w, \underline{\mathfrak{sig}}_\Xi(\,\cdot\,)\big\rangle \cdot w, \tag{58}$$

of $\mathfrak{L}^2_{X_\Xi}$-parametrised ($\mathbb{P}_X$-a.e.-defined) 'simple' functions. Then, the family of random variables

$$\Psi_X(X) \coloneqq \{\psi(X) \mid \psi \in \Psi_X\} \quad \textit{is an } \|\cdot\|_{L^2(\mathcal{H}_\mathcal{Y})}\textit{-dense subset of } L^2_X(\mathcal{H}_\mathcal{Y}). \tag{59}$$

Proof. We first show $\Psi_X(X) \subseteq L^2_X(H_\mathcal{Y})$, for which we fix any $\alpha \in \mathfrak{L}^2_X$. 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. [58, Theorem 1.11], applicable by Remark B.3 (iv)) from the fact that: (a) the compositions $\langle w, \psi_\alpha(X)\rangle = \langle \alpha_w, \underline{\mathfrak{sig}}_\Xi(X)\rangle$ are each Σ_X-measurable by Lemma 3.17 and Lemma 4.3, and (b) $(\langle w, \cdot\rangle \mid w \in [\tilde{d}])$ is a (Schauder) basis of the [topological] dual of $\mathcal{H}_{\tilde{d}}$. (Note that in our setting, Bochner- and Borel-measurability coincide; see Remark 4.1 (ii).)

Next we prove (59), for which we consider any fixed $p \in [1,\infty)$ (a setting that, for the sake of generality, extends beyond the requirements of (59), where only $p = 2$ is needed). For any $\mathbb{Z} \in L^p_X(\mathcal{H}_\mathcal{Y}) =: G$ and any given $w \in [\tilde{d}]^*$, the coordinate $\mathbb{Z}_w := \langle w, \mathbb{Z}\rangle$ is in $L^p(\Omega, \Sigma_X, \mathbb{P}) =: G_1$. (Indeed, $\mathbb{E}\big[|\mathbb{Z}_w|^p\big] = \mathbb{E}\big[\big(|\mathbb{Z}_w|^2\big)^{p/2}\big] \le \mathbb{E}\big[\big(\sum_{w\in[\tilde{d}]^*} \mathbb{Z}_w^2\big)^{p/2}\big] = \mathbb{E}\big[\|\mathbb{Z}\|^p\big] = \|\mathbb{Z}\|^p_{L^p(\mathcal{H}_\mathcal{Y})} < \infty$.)

Now for the $\mathbb{X}^\Xi$-spanned algebra $\mathcal{A}_\Xi(X) := \{\xi(X) \mid \xi \in \mathcal{A}_\Xi\}$, recalling (50) for notation, let

$$H \;:=\; \overline{\mathcal{A}_\Xi(X)}^{L^p} \cap L^\infty \qquad \text{and} \qquad \mathfrak{C} \;:=\; \mathcal{A}_\Xi(X).$$

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 $(50)|_{\Lambda=\Xi}$, 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 3.19 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 $\sup_k |\mathbb{z}_k| \le C$ for some $C > 0$ and $\mathbb{z} := \lim_{k\to\infty} \mathbb{z}_k$ pointwise, we have $\mathbb{z}_k \to \mathbb{z}$ in L^p 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 $\mathfrak{C} = \operatorname{span}_\mathbb{R}\big\{\underline{\xi}^\Xi_w(X) \;\big|\; w \in [\tilde{d}]\big\}$ and hence $\sigma(\mathfrak{C}) = \sigma(\underline{\xi}^\Xi_w(X) \;\big|\; w \in [\tilde{d}])$, Lemma 4.3 implies $\sigma(\mathfrak{C}) = \Sigma_X$, which shows that in fact we proved

$$G_1 \cap L^\infty \;\subseteq\; H \;\subseteq\; \overline{\mathfrak{C}}^{L^p}. \tag{60}$$

Before using this observation to prove (59), note that for the above w-coordinate $\mathbb{Z}_w$ we have

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

(the L^p-convergence holds by dominated convergence). But since $(\mathbb{Z}_w^{\langle n\rangle}) \subset \overline{\mathfrak{C}}^{L^p}$ by (60), we find

$$\mathbb{Z}_w \;\in\; \overline{\mathfrak{C}}^{L^p}, \quad \text{that is:} \quad \mathbb{Z}_w \;=\; L^p\text{-}\lim_{j\to\infty} \langle \alpha_{w,j}, \underline{\mathfrak{sig}}_\Xi(X)\rangle \quad \text{for some } (\alpha_{w,j})_j \subset \mathbb{R}[\underline{d}]. \tag{61}$$

Since (61) holds for $p = 2$ and for all $w \in [\tilde{d}]^*$, the conclusion (59) is now within very close reach:

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

$$\alpha^\star_w \;\in\; \mathbb{R}[\underline{d}] \quad \text{such that} \quad \|\mathbb{Z}_w - \varphi_{\alpha^\star_w}\|^2_{G_1} \;\le\; \varepsilon^2 (2d_\mathcal{Y} + 2)^{-|w|}/2 \qquad \big(w \in [\tilde{d}]^*\big). \tag{62}$$

For the coefficient vector $\alpha^\star := (\alpha^\star_w)_{w\in[\tilde{d}]^*}$, we then obtain

$$\|\alpha^\star\|^2_{\mathfrak{L}} = \sum_{w\in[\tilde{d}]^*} \mathbb{E}\big[\varphi^2_{\alpha^\star_w}\big] \;\le\; 2\sum_{m=0}^\infty \sum_{|w|=m} \|\mathbb{Z}_w - \varphi_{\alpha^\star_w}\|^2_{G_1} + \|\mathbb{Z}_w\|^2_{G_1} \;\le\; \varepsilon^2 \sum_{m=0}^\infty 2^{-m} + 2\|\mathbb{Z}\|^2_G < \infty, \tag{63}$$

where the penultimate inequality is due to there being $\sharp\{w \in [\tilde{d}]^* \mid |w| = m\} = (d_{\mathcal{Y}}+1)^m$ many length-m words in the monoid $[\tilde{d}]^*$. So $\alpha^\star \in \mathfrak{L}^2_{X_\Xi}$ and thus $\psi_{\alpha^\star} \in \Psi_X(X)$, and from (63) we get

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

Since both $\mathbb{Z} \in L^2_X(\mathcal{H}_\mathcal{Y})$ and $\varepsilon > 0$ were arbitrary, the claim (59) is established. □

The next statement, which is of independent interest, is an immediate corollary of the above proof.

Corollary 4.6. *In the setting and notation of Proposition 3.19, we have for any $\upsilon \in \mathcal{M}_1(\mathcal{Z})$ that*

$$\mathcal{A}_\Lambda \;\text{ is an } \|\cdot\|_{L^p(\upsilon)}\text{-dense subset of } L^p(\upsilon), \quad \text{for each } p \in [1,\infty).$$

Proof. This is precisely statement (61), upon replacing $(\Xi, \mathbb{P}_X)$ by (Λ, υ) (which does not impact the proof). □

Corollary 4.7. *Let $k \in \mathbb{N}$ and $\upsilon \in \mathcal{M}_1(\mathcal{X})$, and $\Xi : \mathcal{H}^\downarrow_\mathcal{X} \to \mathcal{H}_\mathcal{X}$ be any feature normalisation. Then for the family of vector-valued (bounded) signature polynomials*

$$\psi^{[k]}_\alpha \;\coloneqq\; \sum_{i=1}^{k} \langle \alpha_i, \underline{\mathfrak{sig}}_\Xi(\cdot)\rangle \cdot e_i \;:\; \mathcal{X} \longrightarrow \mathbb{R}^k, \qquad \alpha \equiv (\alpha_1, \ldots, \alpha_k) \in \mathfrak{L}^2_k \coloneqq \big(\mathbb{R}[\underline{d}]^*\big)^{\times k},$$

we have that $\{\psi^{[k]}_\alpha \,|\, \alpha \in \mathfrak{L}^2_k\}$ is an $\|\cdot\|_{L^2(\upsilon;\mathbb{R}^k)}$-dense subset of $L^2(\upsilon;\mathbb{R}^k)$.

Proof. Immediate by Corollary 4.6 upon recalling the definition in (41) of $\langle\cdot,\cdot\rangle \equiv \langle\cdot,\cdot\rangle_{\mathcal{H}_\mathcal{X}}$. □

4.2 The Conditional Expected Signature of Y given X

Merging Proposition 4.5 with the classical perspective on conditional expectations as L^2-projections (Lemma 2.9) allows us to compute the conditional expected signature $\mathbb{E}[\mathbb{Y}^\Lambda \,|\, X]$ as the solution to a conveniently posed convex optimization problem, as Theorem 4.8 below shows.

In order to align this result with the overarching goal of approximating the regression operator (14) on time-dependent data (51) (Section 2.3), let us recall that, in the notation of Proposition 2.4:

$$\mathbb{E}[\mathbb{Y}^\Lambda \,|\, X] \;=\; \mu.(q) \qquad \text{for} \qquad \mu \coloneqq \mathbb{P}_{(X,Y)} \quad \text{and} \quad q \coloneqq \underline{\mathfrak{sig}}_\Lambda. \tag{64}$$

The following theorem thus provides a principled approximation of the inner component (16) for the proposed regression strategy that combines (17) and (64).

Theorem 4.8. *Adopting the setting and notation of Proposition 4.5, we have that*

$$\mathbb{E}\big[\mathbb{Y}^\Lambda \,\big|\, X\big] \;=\; \lim_{k\to\infty} \psi_{\alpha_k}(X) \quad \textit{in} \quad L^2_X(\mathcal{H}_\mathcal{Y}) \tag{65}$$

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

$$\inf_{\alpha \,\in\, \mathfrak{L}^2_{X_\Xi}} \mathbb{E}\big[\|\mathbb{Y}^\Lambda - \psi_\alpha(X)\|^2\big]. \tag{66}$$

Denoting $\Phi(\mathbb{Z}) \coloneqq \mathbb{E}[\|\mathbb{Y}^\Lambda - \mathbb{Z}\|^2]$ and $\gamma \coloneqq \inf_{\alpha\in\mathfrak{L}^2_{X_\Xi}} \Phi(\psi_\alpha(X))$, the convergence (65) holds $\mathbb{P}$-almost surely if (α_k) runs 'fast enough' in the sense that it satisfies $\sum_{k=0}^\infty (\Phi(\psi_{\alpha_k}(X)) - \gamma)^{1/2} < \infty$.

Sketch of Proof. We combine Proposition 4.5 with an $\mathcal{H}_\mathcal{Y}$-valued version of Lemma 2.9:

Since $\mathbb{Y}_X^\Lambda := \mathbb{E}[\mathbb{Y}^\Lambda \,|\, X]$ is the orthogonal projection of $\underline{\mathfrak{sig}}_\Lambda(Y)$ with respect to the decomposition

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

cf. (Rem. 4.1 (i) and) Lemma 2.9, the Hilbert projection theorem yields the variational characterization

$$\mathbb{Y}_X^\Lambda \;=\; \operatorname*{arg\,min}_{\mathbb{Z}\,\in\, L^2_X(\mathcal{H}_\mathcal{Y})} \mathbb{E}\big\|\mathbb{Y}^\Lambda - \mathbb{Z}\big\|^2, \qquad \text{and hence} \quad \Phi(\mathbb{Y}_X^\Lambda) \;=\; \inf_{\alpha\,\in\,\mathfrak{L}^2_{X_\boxplus}} \Phi(\psi_\alpha(X)), \tag{67}$$

of $\mathbb{Y}_X^\Lambda$ as the L^2-proximum of $\mathbb{Y}^\Lambda$ within the closed subspace $L^2_X(\mathcal{H}_\mathcal{Y})$. Combining this characterisation with Proposition 4.5 leads to the second identity in (67), which in turn implies the desired convergence (65), including the stated almost sure version, via the completeness of $L^2_X(\mathcal{H}_\mathcal{Y})$ and by applying the parallelogram law for the norm $\|\cdot\|_{L^2_X(\mathcal{H}_\mathcal{Y})}$. See Appendix B.2.9 for the details. □

The next result, which is a corollary to the proof of Theorem 4.8, presents a way to directly compute the conditional expectation (20) for a *fixed* function $f \in \mathcal{L}^2(\mathbb{P}_Y; \mathbb{R}^k)$. Ultimately, this computation can be achieved from samples of the pair $(X, Z) := (X, f(Y))$, see Theorem 5.9.

Corollary 4.9. *Let $k \in \mathbb{N}$. For any random vector $Z \in L^2(\Omega, \mathscr{F}, \mathbb{P}; \mathbb{R}^k)$ we have that*

$$\mathbb{E}\big[Z \,\big|\, X\big] \;=\; \lim_{j\to\infty} \psi^{[k]}_{\alpha_j}(X) \quad \textit{in} \quad L^2_X(\mathbb{R}^k) \tag{68}$$

for each minimizing sequence $(\alpha_j) \subset \mathfrak{L}^2_k$ of the convex semi-infinite linear least-squares problem

$$\inf_{\alpha\,\in\,\mathfrak{L}^2_k} \mathbb{E}\Big[\big|Z - \psi^{[k]}_\alpha(X)\big|^2\Big]. \tag{69}$$

Denoting $\Phi_k(\mathcal{W}) := \mathbb{E}[\|Z - \mathcal{W}\|^2]$ and $\eta := \inf_{\alpha\in\mathfrak{L}^2_k} \Phi_m(\psi^{[k]}_\alpha(X))$, the convergence (68) holds $\mathbb{P}$-a.s. if the sequence (α_j) runs 'fast enough' in the sense that it satisfies $\sum_{j=0}^\infty \big(\Phi_k(\psi^{[k]}_{\alpha_j}(X)) - \eta\big)^{1/2} < \infty$.

Proof. The proof of Theorem 4.8 stays valid up to the first identity in (67) if we replace $(\mathbb{Y}^\Lambda, L^2_X(\mathcal{H}_\mathcal{Y}))$ with $\big(Z, L^2_X(\mathbb{R}^k)\big)$, where $L^2_X(\mathbb{R}^k) \equiv \{\mathcal{W} \in L^2(\mathbb{R}^k) \mid \mathcal{W} \text{ is } (\Sigma_X, \mathcal{B}(\mathbb{R}^k))\text{-measurable}\}$. In particular

$$\mathbb{E}\big[Z \,\big|\, X\big] \;=\; \operatorname*{arg\,min}_{\mathcal{W}\,\in\, L^2_X(\mathbb{R}^k)} \mathbb{E}\big|Z - \mathcal{W}\big|^2, \tag{70}$$

and since $\{\psi^{[m]}_\alpha(X) \mid \alpha \in \mathfrak{L}^2_k\}$ is $\|\cdot\|_{L^2_X(\mathbb{R}^k)}$-dense in $L^2_X(\mathbb{R}^k)$ by Corollary 4.7, we see Corollary 4.9 follow from (70) in the same way Theorem 4.8 follows from the first identity in (67). □

Provided that (samples of) the association $(X, Z) := (X, f(Y))$ can be directly observed, the approach of Corollary 4.9 offers a more direct alternative for computing $\mathbb{E}[f(Y) \,|\, X]$ than the compound strategy (17) based on Theorem 4.8 (via (64)). However, there are scenarios where the latter method proves advantageous, such as when the pairs (X, Y) and $(Y, f(Y))$ are observed independently and the evaluation $Y \mapsto f(Y)$ is challenging or infeasible. The compound strategy (17) is also beneficial in cases requiring rapid online evaluations of the regression operator (14) across a wide range of different argument functions (e.g. with LLMs, where individual f may be indicators over the model's vocabulary, cf. (21) and Example 2.8). A further comparison of the advantages and disadvantages of each of these estimation methods is given in Remark B.8.

4.3 Signature-Based Regression of Stochastic Processes

The conditional expectation $\mathbb{E}\big[\mathbb{Y}^\Lambda \,\big|\, X\big]$ of the bounded signature coordinates $\mathbb{Y}^\Lambda$ (see (53)) of a process Y given another process X can be variationally characterized as the solution to a convex least squares problem, see (65) and (66) in Theorem 4.8. This characterisation can be extended to a convex nonparametric estimation method for stochastic process regression (Remark 2.7) through the operator-learning scheme defined by (17) and (64). The central objective of this regression method is the consistent and efficient approximation of X-conditional functionals of Y, see (20), and this subsection presents the probabilistic foundations of this regression approach.

Note that, since the conditional expectation operators in (20) are linear in f, the following results all trivially extend to vector-valued functions of Y.

Proposition 4.10. *For any stochastic processes X and Y as in (52) and any $f \in L^2(\mathbb{P}_Y)$, we have*

$$\mathbb{E}\big[f(Y)\,\big|\,X\big] \;=\; \lim_{l\to\infty} \big\langle \mathbb{E}\big[\mathbb{Y}^\Lambda \,\big|\, X\big],\, \ell_f^{(l)} \big\rangle \quad \textit{in} \quad L^2_X(\mathbb{R}) \tag{71}$$

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

$$\inf_{\ell \in \mathbb{R}[\tilde{d}]} \int_{\mathcal{Y}} \Big(f(y) - \big\langle \ell, \underline{\mathfrak{sig}}_\Lambda(y) \big\rangle \Big)^2 \, \mathbb{P}_Y(\mathrm{d}y). \tag{72}$$

In particular, for $(\ell_f^{(l)})$ as above and any sequence $(\psi_{\alpha_k}(X))$ as in (65) or (168), we have

$$\mathbb{E}\big[f(Y)\,\big|\,X\big] \;=\; \lim_{l\to\infty} \lim_{k\to\infty} \big\langle \psi_{\alpha_k}(X),\, \ell_f^{(l)} \big\rangle \quad \textit{in} \quad L^2_X(\mathbb{R}). \tag{73}$$

Proof. Fix any $f \in L^2(\mathbb{P}_Y)$. By virtue of Corollary 4.6, there is a sequence $(\ell_l) \subset \mathbb{R}[\tilde{d}]$ such that $\int_{\mathcal{Y}} (f(y) - \langle \ell_l, \underline{\mathfrak{sig}}_\Lambda(y)\rangle)^2 \, \mathbb{P}_Y(\mathrm{d}y) = \| f(Y) - \langle \ell_l, \underline{\mathfrak{sig}}_\Lambda(Y)\rangle \|^2_{L^2(\Sigma_Y)} \to 0$ as $l \to \infty$, so the infimum (72) is zero. Hence for any minimizing sequence $(\ell_f^{(l)}) \subset \mathbb{R}[\tilde{d}]$ of (72),

$$\begin{aligned} \big\| \mathbb{E}[f(Y)\,|\,X] - \mathbb{E}[\langle \ell_f^{(l)}, \mathbb{Y}^\Lambda\rangle \,|\, X] \big\|_{L^2(\Sigma_X)} \;&=\; \big\| \mathbb{E}[f(Y) - \langle \ell_f^{(l)}, \mathbb{Y}^\Lambda\rangle \,|\, X] \big\|_{L^2(\Sigma_X)} \\ &\le\; \big\| f(Y) - \langle \ell_f^{(l)}, \mathbb{Y}^\Lambda\rangle \big\|_{L^2(\Sigma_Y)} \longrightarrow\; 0 \quad \text{as } l \to \infty, \end{aligned} \tag{74}$$

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

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

$$\mathbb{E}[\langle \ell_f^{(l)}, \mathbb{Y}^\Lambda\rangle \,|\, X] \;=\; \big\langle \ell_f^{(l)},\, \mathbb{E}[\mathbb{Y}^\Lambda \,|\, X] \big\rangle \qquad (\forall\, l \in \mathbb{N}) \tag{75}$$

with probability one, see for instance [48, Theorem II.2.3]. Combining (74) and (75) proves (71).

The convergence (73) is clear from (71) and (65). Indeed: Since, for each $\ell \in \mathbb{R}[\tilde{d}]$, $\eta_k(\ell) := \langle \psi_{\alpha_k}(X), \ell\rangle \in L^2_X(\mathbb{R})$ is merely a (finite) linear combination of [$L^2_X(\mathbb{R})$-valued] coefficients of $\psi_{\alpha_k}(X)$, the $L^2_X(\mathbb{R})$-convergence $\langle \mathbb{E}[\mathbb{Y}^\Lambda \,|\, X], \ell\rangle = \lim_{k\to\infty} \eta_k(\ell)$ is readily implied by (65) [cf. (54)]. □

Almost-sure versions of (71) and (73) hold, e.g., if Y is compactly supported and f is continuous.

Corollary 4.11. *Let X and Y be as in (52) but with $\mathfrak{D}_Y := \mathrm{supp}(\mathbb{P}_Y)$ compact. Then for any function $f \in C(\mathfrak{D}_Y)$ and any minimizing sequence $(\ell_f^{(l)})_l$ of the optimisation problem*

$$\inf_{\ell \in \mathbb{R}[\tilde{d}]} \big\| f - \langle \ell, \underline{\mathfrak{sig}}_\Lambda(\cdot)\rangle \big\|_{\infty;\mathfrak{D}_Y} \tag{76}$$

and any approximating sequence $(\psi_{\alpha_k}(X))$ *as in* (65) *or* (168)*, we have that*

$$\lim_{l\to\infty} \big\langle \mathbb{E}\big[\mathbb{Y}^\Lambda \,\big|\, X\big],\, \ell_f^{(l)} \big\rangle \overset{\alpha}{=} \mathbb{E}\big[f(Y)\,\big|\,X\big] \overset{\beta}{=} \lim_{l\to\infty}\lim_{k\to\infty} \big\langle \psi_{\alpha_k}(X),\, \ell_f^{(l)} \big\rangle, \tag{77}$$

where the convergence in (α) *holds* $\mathbb{P}$*-a.s. and the convergence in* (β) *holds wrt.* $\|\cdot\|_{L^2_X(\mathbb{R})}$*; if in addition* $(\psi_{\alpha_k}(X))$ *is such that* (65) *or* (169) *converges almost surely, then* (β) *also holds* $\mathbb{P}$*-a.s.*

Proof. See Appendix B.2.10. □

Applied to the special cases $f = \mathbb{1}_A$ for arbitrary Borel sets $A \subseteq \mathcal{Y}$, Proposition 4.10 yields the following variational representation of the conditional law $\mathbb{P}(Y \in \cdot\,|X)$.

Corollary 4.12. *For any stochastic processes* X *and* Y *as in* (52) *and any* $A \in \mathcal{B}(\mathcal{Y})$*, we have*

$$\mathbb{P}(Y \in A \mid X) \;=\; \lim_{l\to\infty}\lim_{k\to\infty} \big\langle \tilde{\psi}_{\alpha_k}(X),\, \ell_A^{(l)} \big\rangle \quad \text{in} \quad L^2_X(\mathbb{R})$$

for each sequence $(\tilde{\psi}_{\alpha_k}(X))$ *as in* (65) *or* (168) *and any minimizing sequence* $\big(\ell_A^{(l)} \,|\, l \in \mathbb{N}\big)$ *of*

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

For an almost sure variant of this result, note that if $A \subseteq \mathcal{Y}$ is open, then $\mathbb{1}_A$ admits a monotone pointwise approximation by a sequence of nonnegative uniformly bounded functions in $C_b(\mathcal{Y})$:

$$\mathbb{1}_A \uparrow h_A^{(\nu)} \quad (\nu\to\infty) \quad \text{pointwise}, \quad \text{e.g.} \quad h_A^{(\nu)} : y \mapsto \frac{\mathrm{d}(y, \mathcal{Y}\backslash A)}{\mathrm{d}(y,\mathcal{Y}\backslash A) + \mathrm{d}(y, F_\nu(A))} \in C_b(\mathcal{Y}), \tag{78}$$

where $\mathrm{d}(y,C) \coloneqq \inf_{z\in C}\|y-z\|_{1\text{-var}}$ (for $C \subset \mathcal{Y}$) and $F_\nu(A) \coloneqq \{y \in \mathcal{Y} \mid \mathrm{d}(y,\mathcal{Y}\backslash A) \geq \nu^{-1}\}$.

Corollary 4.13. *Adopt the setting and notation of Corollary 4.13 and suppose in addition that the support* $\mathfrak{D}_Y \coloneqq \mathrm{supp}(\mathbb{P}_Y)$ *is compact and* $A \subseteq \mathcal{Y}$ *is open. Then*

$$\lim_{\nu\to\infty}\lim_{\mu\to\infty} \big\langle \mathbb{E}\big[\mathbb{Y}^\Lambda \,\big|\, X\big],\, \ell_{A_\nu}^{(\mu)} \big\rangle \overset{\alpha}{=} \mathbb{P}(Y\in A \mid X) \overset{\beta}{=} \lim_{\nu\to\infty}\lim_{\mu\to\infty}\lim_{k\to\infty} \big\langle \tilde{\psi}_{\alpha_k}(X),\, \ell_{A_\nu}^{(\mu)} \big\rangle,$$

which, for any sequence $\big(h_A^{(\nu)}\big)_\nu \subset C_b(\mathcal{Y})$ *as in* (78)*, holds for any* $\big(\ell_{A_\nu}^{(\mu)} \,\big|\, \mu,\nu\in\mathbb{N}\big)$ *such that:*

$$\big(\ell_{A_\nu}^{(\mu)}\big)_\mu \quad \textit{is minimizing sequence of} \quad \inf_{\ell\,\in\,\mathbb{R}[\tilde{d}]} \big\|h_A^{(\nu)} - \langle \ell, \underline{\mathfrak{sig}}_\Lambda(\cdot)\rangle\big\|_{\infty;\mathfrak{D}_Y}, \quad \textit{for each} \quad \nu\in\mathbb{N};$$

under these assumptions, the convergence (α) *holds* $\mathbb{P}$*-a.s. and* (β) *holds wrt.* $L^2_X(\mathbb{R})$*-convergence, and* (β) *holds* $\mathbb{P}$*-a.s. if* $(\tilde{\psi}_{\alpha_k}(X))$ *is chosen such that* (65) *or* (168) *converges almost surely.*

Since $\mathbb{P}(Y\in A \mid X) = 1 - \mathbb{P}(Y \in A^c \mid X)$, analogous statements hold if A is closed.

Proof. From (78) and the definition of conditional probability, we have

$$\mathbb{P}(Y\in A\mid X) \;=\; \mathbb{E}\big[\mathbb{1}_A(Y)\,\big|\,X\big] \;=\; \lim_{\nu\to\infty} \mathbb{E}\big[h_A^{(\nu)}(Y)\,\big|\,X\big] \quad \mathbb{P}\text{-a.s.} \tag{79}$$

via the conditional monotone convergence theorem. The remainder of the corollary then follows from (79) by applying Proposition 4.10 to the continuous functions $f = h_A^{(\nu)}\big|_{\mathfrak{D}_Y}$ for each $\nu\in\mathbb{N}$. □

5 Signature-Based Regression: Estimation and Consistency

This section extends the probabilistic results of Section 4.3 to a practically applicable, convex, and nonparametric statistical method to perform regression on and between stochastic processes. Specifically, we derive two signature-based regression estimators (Algorithms I and II, cf. Definition 2.6) whose universal consistency guarantees come with explicit convergence rates (Theorem 5.9).

These estimators rely on a data-based discretization of the optimisation problems (66) and (69) and of the associated approximations (72) and (73). The initial hurdle is that the regressors $\Psi_X \equiv \{\psi_\alpha \,|\, \alpha \in \mathfrak{L}_X^2\}$, while defined only almost everywhere wrt. $\mathbb{P}_X$ in (58), require pointwise evaluation when applied to specific data points. This is addressed in Section 5.1 by embedding Proposition 4.5 within the context of vector-valued reproducing kernel Hilbert spaces. In this framework, Section 5.2 adapts the tools and ideas from Section 4 to prepare for a regularised-least-squares-type empirical estimation analysis. This leads to several auxiliary results that structure the algorithmisation of Proposition 4.10 and culminate in Section 5.3, where these notions are combined to a first estimator for the regression operator $f \mapsto \mu_\cdot(f)$ in (20). Section 5.4 specialises this approach to finite-dimensional regressands (Corollary 4.9) and proposes a second, specialised estimator for this scenario. Finally, Section 5.5 synthesizes all previous ideas and results to demonstrate the statistical consistency of the proposed estimators, convergence rates and error bounds (Theorem 5.9).

5.1 Signature Regressors form a Dense Vector-Valued RKHS

To underpin the approximation strategies of Section 4 with a clean and robust estimation theory, it is beneficial to transfer the insights of Proposition 4.5 to the framework of vector-valued reproducing kernel Hilbert spaces (vRKHS), cf. [42]. To this end, we modify the function space in (58) to obtain a dense subsystem of functions in $L^2_X(\mathcal{H}_\mathcal{Y})$ with vRKHS structure, thus ensuring that these modified signature-regressors can be evaluated pointwise. This adaptation is crucial for empirically approximating (via (68) and (71)) the probabilistic quantities in (13) from discrete-data points.

(The theory of vRKHS was first developed in [52], and we refer to [7] for a modern exposition.)

Proposition 5.1. *Let Ξ be a feature normalisation on $\mathcal{H}_\mathcal{X}$, and denote by $(\mathcal{H}_\kappa, \|\cdot\|_\kappa) \subset \mathbb{R}^\mathcal{X}$ the RKHS with reproducing kernel $\kappa : (x,y) \mapsto \langle \underline{\mathfrak{sig}}_\Xi(x), \underline{\mathfrak{sig}}_\Xi(y)\rangle$. Then the space of functions*

$$\mathfrak{H}_\Xi \coloneqq \Big\{ f : \mathcal{X} \to \mathcal{H}_\mathcal{Y} \;\Big|\; \{\langle f, w\rangle \,|\, w \in [\tilde{d}]^*\} \subset \mathcal{H}_\kappa \;\; \text{and} \;\; \textstyle\sum_{w\in[\tilde{d}]^*} \|\langle f, w\rangle\|_\kappa^2 < \infty \Big\} \tag{80}$$

is a separable vRKHS with reproducing kernel $K : \mathcal{X}^{\times 2} \to \mathfrak{L}(\mathcal{H}_\mathcal{Y}), (x,y) \mapsto \kappa(x,y)\mathrm{Id}_{\mathcal{H}_\mathcal{Y}}$. Moreover,

$$\mathfrak{H}_\Xi \quad \text{is dense in} \quad \big(L^2(\mathcal{X}, \mathbb{P}_X; \mathcal{H}_\mathcal{Y}), \|\cdot\|_{L^2(\mathbb{P}_X;\mathcal{H}_\mathcal{Y})}\big), \tag{81}$$

and $\mathfrak{H}_\Xi^{[m]} = \{\tilde{f} \equiv \pi_{[m]} \circ f \mid f \in \mathfrak{H}_\Xi\}\big)$ is dense in $\big(L^2(\mathcal{X}, \mathbb{P}_X; \pi_{[m]}(\mathcal{H}_\mathcal{Y})), \|\cdot\|_{L^2(\mathbb{P}_\mathcal{X};\pi_{[m]}(\mathcal{H}_\mathcal{Y}))}\big)$ for each $m \geq 0$, where $\mathfrak{H}_\Xi^{[0]} \cong \mathcal{H}_\kappa$. Recalling Proposition 4.5 and introducing the space

$$\Psi_X^\star \coloneqq \Big\{ \textstyle\sum_{w\in[\tilde{d}]^*} \langle u_w, \underline{\mathfrak{sig}}_\Xi(\cdot)\rangle w \;\Big|\; (u_w) \subset \mathcal{H}_\mathcal{X} \,:\, \sum_{w\in[\tilde{d}]^*} \mathbb{E}\big[\langle u_w, \mathbb{X}^\Xi\rangle^2\big] < \infty \Big\},$$

we further have the $\|\cdot\|_{L^2(\mathcal{X},\mathbb{P}_X;\mathcal{H}_\mathcal{Y})}$-dense—and, in the case of $\mathfrak{H}_\Xi$, also continuous—embeddings

$$\mathfrak{H}_\Xi \,, \Psi_X \;\subseteq\; \Psi_X^\star \subseteq\; L^2(\mathcal{X}, \mathbb{P}_X; \mathcal{H}_\mathcal{Y}). \tag{82}$$

Proof. See Appendix B.3.1. □

5.2 Preliminary Results via Hilbert-Valued Regularised Least Squares

Proposition 5.1 together with Proposition 4.10 and Corollary 4.9, allows us to leverage the consistency theory of 'vectorial' regularised least squares routines, as outlined in [42], to derive estimators that are universally consistent in the sense of Definition 2.6. This section provides some auxiliary results to support this.

For this, we let $\mu \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ be a Borel probability measure with marginals $\mu_{\mathcal{X}} \coloneqq \mu \circ \pi_{\mathcal{X}}^{-1}$ and $\mu_{\mathcal{Y}} \coloneqq \mu \circ \pi_{\mathcal{Y}}^{-1}$, and appeal to an M-estimation scheme given by the following family of objectives:

$$\begin{aligned}
&\boldsymbol{\phi}_{\mu}^{(m)} \,:\, L^2\big(\mu_{\mathcal{X}}; \mathcal{H}_{\mathcal{Y}}^{[m]}\big) \longrightarrow \overline{\mathbb{R}}_+, \quad g \mapsto \boldsymbol{\phi}_{\mu}^{(m)}(g) \coloneqq \int_{\mathcal{X}\times\mathcal{Y}} \big\| \pi_{[m]}(\underline{\mathfrak{sig}}_{\Lambda}(y)) - g(x) \big\|^2 \,\mathrm{d}\mu(x,y), \quad (m \in \overline{\mathbb{N}}),\\
&\boldsymbol{\phi}_{\mu|\lambda}^{(m)} \,:\, \mathfrak{H}_{\Xi}^{[m]} \longrightarrow \overline{\mathbb{R}}_+, \qquad\qquad g \mapsto \boldsymbol{\phi}_{\mu|\lambda}^{(m)}(g) \coloneqq \boldsymbol{\phi}_{\mu}^{(m)}(g) + \lambda \|g\|_{\mathfrak{H}_{\Xi}}^2, \qquad\qquad (\lambda > 0),
\end{aligned}$$

where $\mathcal{H}_{\mathcal{Y}}^{[m]} \coloneqq \pi_{[m]}\big(\mathcal{H}_{\mathcal{Y}}\big)$ and $\mathfrak{H}_{\Xi}^{[m]} \coloneqq \pi_{[m]}\big(\mathfrak{H}_{\Xi}\big)$ for $\pi_{[m]}$ as in Section 3.10 and $\pi_{[\infty]} \coloneqq \mathrm{id}_{\mathcal{H}_{\mathcal{Y}}}$.

For any given $f \in \mathcal{L}^2(\mu_{\mathcal{Y}})$, we additionally consider

$$\begin{aligned}
&\boldsymbol{\phi}_{\mu_{\mathcal{Y}}}^{f} \,:\, \mathcal{H}_{\mathcal{Y}} \longrightarrow \mathbb{R}_+, \quad \ell \mapsto \boldsymbol{\phi}_{\mu_{\mathcal{Y}}}^{f}(\ell) \coloneqq \int_{\mathcal{Y}} \big| f(y) - \langle \ell, \underline{\mathfrak{sig}}_{\Lambda}(y) \rangle \big|^2 \mu_{\mathcal{Y}}(\mathrm{d}y),\\
&\boldsymbol{\phi}_{\mu_{\mathcal{Y}}|\lambda}^{f} \,:\, \mathcal{H}_{\mathcal{Y}} \longrightarrow \mathbb{R}_+, \quad \ell \mapsto \boldsymbol{\phi}_{\mu_{\mathcal{Y}}|\lambda}^{f}(\ell) \coloneqq \boldsymbol{\phi}_{\mu_{\mathcal{Y}}}^{f}(\ell) + \lambda \big\| \langle \ell, \underline{\mathfrak{sig}}_{\Lambda}(\cdot) \rangle \big\|_{\kappa}^2.
\end{aligned} \tag{83}$$

Finally, for non-zero $(f, \Upsilon) \in L^2(\mu_{\mathcal{Y}}) \times L^2(\mu_{\mathcal{X}}; \mathcal{H}_{\mathcal{Y}})$ we define

$$\vartheta_f(\tilde{\epsilon}) \coloneqq \frac{\tilde{\epsilon}^2}{4}\Big(\inf\big\{ \|g\|_{\kappa}^2 \,\big|\, g \in \mathcal{H}_{\kappa} :\, \|f - g\|_{L^2(\mathbb{P}_{\mathcal{Y}})}^2 \le \tilde{\varepsilon}^2/2 \big\} \Big)^{-1} \quad \text{and} \tag{84}$$

$$\tilde{\vartheta}_{\Upsilon}(\tilde{\epsilon}) \coloneqq \frac{\tilde{\epsilon}^2}{4}\Big(\inf\big\{ \|g\|_{\mathfrak{H}_{\Xi}}^2 \,\big|\, g \in \mathfrak{H}_{\Xi}^{[m]} :\, \|\Upsilon - g\|_{L^2(\mathbb{P}_{\mathcal{X}}; \mathcal{H}_{\mathcal{Y}})}^2 \le \tilde{\varepsilon}^2/2 \big\} \Big)^{-1}, \tag{85}$$

where we set $\vartheta_f(\tilde{\epsilon}) \coloneqq \infty$ [resp. $\tilde{\vartheta}_{\Upsilon}(\tilde{\epsilon}) \coloneqq \infty$] if the infimum in (84) [resp. in (85)] is zero.

(Note $\tilde{\epsilon} \mapsto \vartheta_f(\tilde{\epsilon})$ and $\tilde{\epsilon} \mapsto \tilde{\vartheta}_{\Upsilon}^{(m)}(\tilde{\epsilon})$ are well-defined maps from $(0, \infty)$ to $(0, \infty]$ by Proposition 5.1.)

The following technical results elucidate basic properties and interrelations of the above auxiliary functions, preparing them for use in deriving our announced regression estimators.

Lemma 5.2. *For each $f \in \mathcal{L}^2(\mu_{\mathcal{Y}})$ and any $\lambda > 0$ and $m \in \overline{\mathbb{N}}$, both minimizers*

$$\ell_{\mu_{\mathcal{Y}},\lambda}^{f} \in \operatorname*{arg\,min}_{\ell \,\in\, \mathcal{H}_{\mathcal{Y}}} \boldsymbol{\phi}_{\mu_{\mathcal{Y}}|\lambda}^{f}(\ell) \qquad \textit{and} \qquad \Upsilon_{\mu,\lambda,m}^{*} = \operatorname*{arg\,min}_{g \,\in\, \mathfrak{H}_{\Xi}^{[m]}} \boldsymbol{\phi}_{\mu|\lambda}^{(m)}(g) \tag{86}$$

exist and the minimizer $\Upsilon_{\mu,\lambda,m}^{*}$ is unique. Moreover, if $\mu_{\mathcal{Y}}$ is an empirical measure of the form

$$\mu_{\mathcal{Y}} = \frac{1}{|I|} \sum_{i \in I} \delta_{y_i} \qquad \textit{for some} \quad \{y_i \mid i \in I\} \subset \mathcal{Y} \quad \textit{with } I \textit{ finite}, \tag{87}$$

then $\big(\arg\min_{\ell \in \mathcal{H}_{\mathcal{Y}}} \boldsymbol{\phi}_{\mu_{\mathcal{Y}}|\lambda}^{f}\big) \cap V_{\mu_{\mathcal{Y}}} \neq \emptyset$ for $V_{\mu_{\mathcal{Y}}} \coloneqq \mathrm{span}\{\underline{\mathfrak{sig}}_{\Lambda}(y_i) \mid i \in I\}$; if, in addition, the points in $\{y_i \mid i \in I\}$ are pairwise distinct, then $\ell_{\mu_{\mathcal{Y}},\lambda}^{f,\star} \coloneqq \arg\min_{\ell \in V_{\mu_{\mathcal{Y}}}} \boldsymbol{\phi}_{\mu_{\mathcal{Y}}|\lambda}^{f}(\ell)$ is unique and satisfies:

$$\big\| \ell_{\mu_{\mathcal{Y}},\lambda}^{f} \big\|_{\mathcal{H}_{\mathcal{Y}}} \ge \big\| \ell_{\mu_{\mathcal{Y}},\lambda}^{f,\star} \big\|_{\mathcal{H}_{\mathcal{Y}}} \quad \textit{for each} \quad \ell_{\mu_{\mathcal{Y}},\lambda}^{f} \textit{ as in (86)}, \tag{88}$$

and also $\big\| \ell_{\mu_{\mathcal{Y}},\lambda}^{f,\star} \big\|_{\mathcal{H}_{\mathcal{Y}}} = \big\| \arg\min_{h \in \mathcal{H}_{\tilde{\kappa}}} \tilde{\phi}_{f,\lambda}(h) \big\|_{\tilde{\kappa}}$ for the map $\tilde{\phi}_{f,\lambda} : \mathcal{H}_{\tilde{\kappa}} \ni h \mapsto \|f - h\|_{L^2(\mu_{\mathcal{Y}})}^2 + \lambda \|h\|_{\tilde{\kappa}}^2$, where the domain $\mathcal{H}_{\tilde{\kappa}}$ is an RKHS with kernel $\tilde{\kappa} : \mathcal{Y}^{\times 2} \ni (x,y) \mapsto \langle \underline{\mathfrak{sig}}_{\Lambda}(x), \underline{\mathfrak{sig}}_{\Lambda}(y) \rangle$.

Proof. Delegated to Appendix B.3.2. □

In the following, suppose that, for any fixed X and Y as in (52), we are given observations

$$\mathfrak{Y} \equiv \{Y^{(l)} \mid l \in [n]\} \quad \text{and} \quad \mathfrak{Z} \equiv \{(X^{(j)}, Y^{(j)}) \mid j \in [N]\} \qquad (n, N \in \mathbb{N}), \tag{89}$$

modelled as iid copies of Y and (X, Y), respectively.

For the associated empirical (random) measures $\hat{\mu}_{\mathfrak{Y}} \coloneqq \frac{1}{n}\sum_{l=1}^{n} \delta_{Y^{(l)}}$ and $\hat{\mu}_{\mathfrak{Z}} \coloneqq \frac{1}{N}\sum_{j=1}^{N} \delta_{(X^{(j)},Y^{(j)})}$ and for any given $f \in \mathcal{L}^2(\mathbb{P}_Y)$, we follow the notation in (86) and choose any minimizers

$$\hat{\ell}^f_{n,\lambda} \coloneqq \ell^f_{\hat{\mu}_{\mathfrak{Y}},\lambda} \qquad \text{and} \qquad \hat{\Upsilon}^*_{N,\lambda,m} \coloneqq \Upsilon^*_{\hat{\mu}_{\mathfrak{Z}},\lambda,m} \qquad \text{and} \qquad \hat{\Upsilon}^*_{N,\lambda} \coloneqq \Upsilon^*_{\hat{\mu}_{\mathfrak{Z}},\lambda,\infty}. \tag{90}$$

Recall that the objects in (90) are all ($\mathfrak{Y}$- resp. $\mathfrak{Z}$-dependent) random variables.

We describe the empirical minimizers in (90) and distinguish three associated types of error.

Proposition 5.3. *Let $\mathfrak{Y}$, $\mathfrak{Z}$ be as in (89), and consider any $f \in \mathcal{L}^2(\mathbb{P}_Y)$. Then for all $\lambda > 0$,*

$$\langle \hat{\Upsilon}^*_{N,\lambda,m}, w\rangle \in \operatorname{span}\big(\langle \underline{\mathfrak{sig}}_{\Xi}(X^{(j)}), \underline{\mathfrak{sig}}_{\Xi}(\cdot)\rangle \,\big|\, j \in [N]\big), \quad \text{for each } \; w \in [\tilde{d}]^* : |w| \le m \quad (m \in \overline{\mathbb{N}}). \tag{91}$$

Moreover, with $m_{\tilde{d}} \coloneqq \sum_{\nu=0}^{m} \tilde{d}^{\nu}$ and $q^*_j \coloneqq \langle \underline{\mathfrak{sig}}_{\Xi}(X^{(j)}), \underline{\mathfrak{sig}}_{\Xi}(\cdot)\rangle$ we have that, with probability one,

$$\sum_{l=1}^{n} \alpha_l \, \underline{\mathfrak{sig}}_{\Lambda}(Y^{(l)}) \;\in\; \operatorname*{arg\,min}_{\ell \in \mathcal{H}_{\mathcal{Y}}} \phi^f_{\hat{\mu}_{\mathfrak{Y}}|\lambda}(\ell) \qquad \text{and} \qquad \hat{\Upsilon}^{\mathrm{lex}}_{N,\lambda,m} = \tilde{A}_*^{\top} \cdot (q^*_j)_{j=1}^{N} \tag{92}$$

if and only if the coefficients $\alpha^* \equiv (\alpha_l)_{l=1}^n \in \mathbb{R}^n$ and $\tilde{A}_* \equiv \big(\tilde{a}_{(1)} \big| \cdots \big| \tilde{a}_{(m_{\tilde{d}})}\big) \in \mathbb{R}^{N \times m_{\tilde{d}}}$ are such that

$$(\boldsymbol{A}_{\mathfrak{Y}}^{\top}\boldsymbol{A}_{\mathfrak{Y}} + n\lambda \boldsymbol{A}_{\mathfrak{Y}})\, \alpha^* = \boldsymbol{A}_{\mathfrak{Y}} b_{\mathfrak{Y},f} \qquad \text{and} \tag{93}$$

$$(\boldsymbol{A}_{\mathfrak{Z}}^{\top}\boldsymbol{A}_{\mathfrak{Z}} + N\lambda \boldsymbol{A}_{\mathfrak{Z}})\tilde{A}_* = \boldsymbol{A}_{\mathfrak{Z}}\boldsymbol{B}_{\mathfrak{Z}}, \tag{94}$$

where the above systems are formulated in terms of the data-based (random) matrices and vectors

$$\boldsymbol{A}_{\mathfrak{Y}} \coloneqq \big(\langle \underline{\mathfrak{sig}}_{\Lambda}(Y^{(k)}), \underline{\mathfrak{sig}}_{\Lambda}(Y^{(l)})\rangle\big)_{k,l=1}^{n} \quad \text{and} \quad b_{\mathfrak{Y},f} \coloneqq \big(f(Y^{(l)})\big)_{l=1}^{n}, \tag{95}$$

$$\boldsymbol{A}_{\mathfrak{Z}} \coloneqq \big(\langle \underline{\mathfrak{sig}}_{\Xi}(X^{(i)}), \underline{\mathfrak{sig}}_{\Xi}(X^{(j)})\rangle\big)_{i,j=1}^{N} \quad \text{and} \quad \boldsymbol{B}_{\mathfrak{Z}} \coloneqq \big(\langle \underline{\mathfrak{sig}}_{\Lambda}(Y^{(i)}), \eta^{-1}(j)\rangle\big)_{(i,j)\in[N]\times[m_{\tilde{d}}]}; \tag{96}$$

here, $\eta : [\tilde{d}]^* \to \mathbb{N}_0$ is the length-lexicographic[7] (or shortlex) ordering of the words in $[\tilde{d}]^*$ and

$$\hat{\Upsilon}^{\mathrm{lex}}_{N,\lambda,m} \equiv \big(\langle \hat{\Upsilon}_{N,\lambda,m}, \eta^{-1}(1)\rangle, \ldots, \langle \hat{\Upsilon}_{N,\lambda,m}, \eta^{-1}(m_{\tilde{d}})\rangle\big) : \mathcal{X} \to \mathbb{R}^{m_{\tilde{d}}} \tag{97}$$

is the length-lexicographically indexed version of $\hat{\Upsilon}^*_{N,\lambda,m}$.

Proof. See Appendix B.3.3. □

Remark 5.4. The matrix $\boldsymbol{A}_{\mathfrak{Z}}$ in (95) [resp. the matrix $\boldsymbol{A}_{\mathfrak{Y}}$ in (96)] is invertible iff the observed samples $(X^{(j)})_j$ in $\mathfrak{Z}$ [resp. the samples $(Y^{(l)})_l$ in $\mathfrak{Y}$] are pairwise disjoint. For a proof, see the argumentation around (151) in Appendix B.3.2. ◆

[7] Recall that in the length-lexicographic ordering of $[\tilde{d}]^*$, words are primarily sorted by wordlength with the shortest words first, and words of the same length are sorted into lexicographical order; see e.g. [56, p. 14].

Revisiting the notation from (86), let us now complement (90) by any fixed choice of minimizers

$$\ell^f_\lambda \coloneqq \ell^f_{\mathbb{P}_Y,\lambda} \qquad \text{and} \qquad \Upsilon^*_{\lambda,m} \coloneqq \Upsilon^*_{\mathbb{P}_{(X,Y)},\lambda,m} \qquad \text{and} \qquad \Upsilon^*_\lambda \coloneqq \Upsilon^*_{\mathbb{P}_{(X,Y)},\lambda,\infty}.$$

To extend the uniqueness and optimality characteristic (88) to singular kernel matrices (95), let

$$\hat{\ell}^\star_{n,\lambda,f} \coloneqq \sum_{l=1}^n \alpha^\star_l \, \underline{\mathfrak{sig}}_\Lambda(Y^{(l)}) \quad \text{for} \quad \alpha^\star \equiv \big(\alpha^\star_l\big)_{l=1}^n \quad \text{the minimum norm solution of (93)}; \tag{98}$$

equivalently, $\alpha^\star = (\boldsymbol{A}_{\mathfrak{Y}}^\top \boldsymbol{A}_{\mathfrak{Y}} + n\lambda \boldsymbol{A}_{\mathfrak{Y}})^+ \boldsymbol{A}_{\mathfrak{Y}} b_{\mathfrak{Y},f}$ for $(\,\cdot\,)^+$ the Moore-Penrose pseudoinverse [43].

Lemma 5.5. *In the above notation, consider for any given $f \in \mathcal{L}^2(\mathbb{P}_Y)$ the expressions*

$$r^{\mathrm{I}}_f(\lambda) \coloneqq \big\| f - \langle \ell^f_\lambda, \underline{\mathfrak{sig}}_\Lambda \rangle \big\|_{L^2(\mathbb{P}_Y)} \quad \textit{and} \quad r^{\mathrm{II}}_{f,\lambda}(n) \coloneqq \big\| \langle \ell^f_\lambda - \hat{\ell}^f_{n,\lambda}, \underline{\mathfrak{sig}}_\Lambda \rangle \big\|_{L^2(\mathbb{P}_Y)}, \tag{99}$$

referred to as the f-approximation error *and the* f-estimation error*, respectively. Set further*

$$r^{\mathrm{III}}_{f,\lambda,n}(m) \coloneqq \big\| \hat{\ell}^\star_{n,\lambda,f} - \pi_{[m]}(\hat{\ell}^\star_{n,\lambda,f}) \big\|_{\mathcal{H}_\mathcal{Y}}, \tag{100}$$

which, for a given $m \in \overline{\mathbb{N}}$, we refer to as the f-cutoff error*. Recalling that $\mathbb{E}[\mathbb{Y}^\Lambda \,|\, X] = \varphi_\star(X)$ for some unique $\varphi_\star \in L^2(\mathbb{P}_X; \mathcal{H}_\mathcal{Y})$ by the Doob-Dynkin lemma, let us finally consider the expressions*

$$R^{\mathrm{I}}_m(\lambda) \coloneqq \big\| \pi_{[m]}\big(\varphi_\star - \Upsilon^*_\lambda\big) \big\|_{L^2(\mathbb{P}_X;\mathcal{H}_\mathcal{Y})} \quad \textit{and} \quad R^{\mathrm{II}}_{m,\lambda}(N) \coloneqq \big\| \pi_{[m]}\big(\Upsilon^*_\lambda - \hat{\Upsilon}^*_{N,\lambda}\big) \big\|_{L^2(\mathbb{P}_X;\mathcal{H}_\mathcal{Y})}, \tag{101}$$

referred to as the inner approximation error *and* inner estimation error*, respectively. Then we have:*

(i) $$\sup\big\{ r^{\mathrm{I}}_f(\lambda) \,\big|\, 0 < \lambda \leq^{8} \vartheta_f(\epsilon) \big\} \;\leq\; \epsilon, \quad \textit{for all} \;\; \epsilon > 0\,;$$

(ii) *it holds that* $\pi_{[m]}(\varphi_\star) = \arg\min_{g \,\in\, L^2(\mathbb{P}_X;\mathcal{H}^{[m]}_\mathcal{Y})} \boldsymbol{\phi}^{(m)}_{\mathbb{P}_{(X,Y)}}(g)$ *for each $m \in \overline{\mathbb{N}}$, and also:*

$$R^{\mathrm{I}}_m(\lambda)^2 \;=\; \boldsymbol{\phi}^{(m)}_{\mathbb{P}_{(X,Y)}}\big(\pi_{[m]}(\Upsilon^*_\lambda)\big) - \boldsymbol{\phi}^{(m)}_{\mathbb{P}_{(X,Y)}}\big(\pi_{[m]}(\varphi_\star)\big), \quad \textit{for all} \;\; \lambda > 0\,;$$

(iii) $$\textit{for each} \;\; m \in \overline{\mathbb{N}} : \quad \sup\big\{ R^{\mathrm{I}}_m(\lambda) \,\big|\, 0 < \lambda \leq \tilde{\vartheta}_{\pi_{[m]}(\varphi_\star)}(\epsilon) \big\} \;\leq\; \epsilon, \quad \textit{for all} \;\; \epsilon > 0\,;$$

(iv) *there is an explicit monotone decreasing null sequence $\big(\beta^{(\mathrm{cut})}_m\big) \subset \mathbb{R}_+$, depending on $(\mathfrak{Y}, \Lambda, \lambda, f)$, such that with probability one:*

$$r^{\mathrm{III}}_{f,\lambda,n}(m) \leq \beta^{(\mathrm{cut})}_m \quad \textit{for all } m \geq 0\,;$$

(v) *for any $\delta > 0$, there are explicit strictly decreasing null sequences $(\beta^{(\mathrm{out})}_n), (\beta^{(\mathrm{in})}_N) \subset \mathbb{R}_+$, depending on $(\Lambda, \lambda, \delta)$ and $(\Lambda, \Xi, \lambda, \delta)$ respectively, such that for each $n, N \in \mathbb{N}$,*

$$\mathbb{P}\big(r^{\mathrm{II}}_{f,\lambda}(n) \geq \beta^{(\mathrm{out})}_n\big) \leq \delta \qquad \textit{and} \qquad \mathbb{P}\big(R^{\mathrm{II}}_{m,\lambda}(N) \geq \beta^{(\mathrm{in})}_N\big) \leq \delta. \tag{102}$$

Proof. Reported in Appendix B.3.4. □

Remark 5.6. Specific rate sequences $(\beta^{(\mathrm{out})}_n)$ and $(\beta^{(\mathrm{in})}_N)$ for which (102) holds are, for example,

$$\beta^{(\mathrm{out})}_n \coloneqq \frac{c_\Lambda}{\lambda\sqrt{n\delta}} \qquad \text{and} \qquad \beta^{(\mathrm{in})}_N \coloneqq \sqrt{\frac{c_\Xi c_\Lambda}{\lambda^2 N \delta}} \qquad (n, N \in \mathbb{N}) \tag{103}$$

with constants $c_Z \coloneqq \sup_{\boldsymbol{t} \in \mathcal{H}_\xi} \|Z(\boldsymbol{t})\|^2_{\mathcal{H}_\xi}$ for $(Z, \xi) \in \{(\Lambda, \mathcal{Y}), (\Xi, \mathcal{X})\}$; see proof of Lemma 5.5 (v). ◆

[8] If $\vartheta_f(\epsilon) = \infty$, then the constraint '$\lambda \leq \vartheta_f(\epsilon)$' is to be dropped; likewise for (iii).

5.3 A Convex Estimator for Stochastic Process Regression

Let now $f = (f_1, \cdots, f_k) \in \mathcal{L}^2(\mathbb{P}_Y; \mathbb{R}^k)$, for any $k \in \mathbb{N}$. In order to structure the results of Sections 5.1 and 5.2 into an estimator for $\mathbb{E}[f(Y) \,|\, X]$, we introduce, for each $\varepsilon, \delta > 0$ and $\lambda > 0$, the moduli

$$\begin{aligned} \lambda^{\mathrm{I}}_f(\varepsilon) &:= \min\{\vartheta_{f_i}(\varepsilon/(5k)) \mid i = 1, \ldots, k\} \quad \text{and} \\ n_{f,\lambda}(\varepsilon, \delta) &:= \min\left\{n \in \mathbb{N} \,\middle|\, \beta^{(\mathrm{out})}_n(\delta/(2k); \lambda) \leq \varepsilon/(5k)\right\} \end{aligned} \tag{104}$$

for any fixed rate sequence $\big(\beta^{(\mathrm{out})}_n\big) \equiv \big(\beta^{(\mathrm{out})}_n(\delta; \lambda)\big)$ as in Lemma 5.5 (v), e.g. the one in (103), and

$$m_{f,\lambda}(\varepsilon; \mathfrak{Y}) := \min\left\{m \in \mathbb{N} \,\middle|\, \|\boldsymbol{\alpha}^f_\lambda\|_{1,2} \tilde{\beta}_m \leq \varepsilon/(5k)\right\} \tag{105}$$

for the norm $\|A\|_{1,2} := \max_{1 \leq j \leq n} \|A_{:j}\|_2$ and the $\mathfrak{Y}$-dependent sequence $(\tilde{\beta}_m)$ from (158), and with

$$\boldsymbol{\alpha}^f_\lambda \equiv \big(\alpha^{(\lambda)}_{li}\big)_{(l,i) \in [n] \times [k]} := \big(\boldsymbol{A}_{\mathfrak{Y}}^\top \boldsymbol{A}_{\mathfrak{Y}} + n\lambda \boldsymbol{A}_{\mathfrak{Y}}\big)^+ \boldsymbol{A}_{\mathfrak{Y}} B_{\mathfrak{Y},f}$$

for the data matrices $\boldsymbol{A}_{\mathfrak{Y}}$ from (95) and $B_{\mathfrak{Y},f} := \big(f_j(Y^{(l)})\big)_{(l,j) \in [n] \times [k]}$. For $m \in \mathbb{N}$, we further set

$$\begin{aligned} \lambda^{\mathrm{I}}_{\lambda,m}(\varepsilon; \mathfrak{Y}) &:= \tilde{\vartheta}_{\varphi^\star_m}\big(\varepsilon/(5c_{\lambda,m;\mathfrak{Y}})\big) \qquad \text{and} \\ N_{\lambda,\tilde{\lambda},m}(\varepsilon, \delta; \mathfrak{Y}) &:= \min\left\{N \in \mathbb{N} \,\middle|\, \beta^{(\mathrm{in})}_N(\delta/2; \tilde{\lambda}) \leq \varepsilon/(5c_{\lambda,m;\mathfrak{Y}})\right\} \end{aligned} \tag{106}$$

for $\varphi^\star_m := \pi_{[m]}(\varphi_\star)$ as in Lemma 5.5, with the constant $c_{\lambda,m;\mathfrak{Y}} := \sqrt{\sum_{i=1}^k \|\pi_{[m]}(\hat{\ell}^\star_{n,\lambda,f_i})\|^2_{\tilde{\mathcal{H}}_{\mathcal{Y}}}}$ computable from (98) and for any fixed sequence $(\beta^{(\mathrm{in})}_N) \equiv (\beta^{(\mathrm{in})}_N(\tilde{\delta}; \tilde{\lambda}))$ as in Lem. 5.5 (v), e.g. (103). Finally, for any dimensions $n, N \in \mathbb{N}$ of the data sets (89) and parameters $\underline{\lambda} \equiv (\lambda_1, \lambda_2) \in \mathbb{R}^2_{>0}$ and $m \in \mathbb{N}$, consider the estimator (of the regression function between $f(Y)$ and X)

$$\hat{\mathscr{R}}^{\mathrm{I}}_{\underline{\lambda},f}[n, N, m] := \sum_{i=1}^k \left[\sum_{l=1}^n \alpha^{\lambda_1}_{li} \big\langle \pi^{\mathrm{lex}}_{[m]}(\underline{\mathfrak{sig}}_\Lambda(Y^{(l)})), \hat{\Upsilon}^{\mathrm{lex}}_{N,\lambda_2,m}(\cdot)\big\rangle_2\right] e_i \,:\, \mathcal{X} \longrightarrow \mathbb{R}^k \tag{107}$$

with $(e_i \mid i \in [k])$ the standard basis of $\mathbb{R}^k$, $\langle \cdot, \cdot \rangle_2$ the Euclidean inner product on $\mathbb{R}^{m_{\tilde{d}}}$, and where

$$\hat{\Upsilon}^{\mathrm{lex}}_{N,\lambda_2,m} = \sum_{\nu=1}^{m_{\tilde{d}}} \left[\sum_{j=1}^N \beta^{(\lambda_2)}_{j\nu} \big\langle \underline{\mathfrak{sig}}_\Xi(X^{(j)}), \underline{\mathfrak{sig}}_\Xi(\cdot)\big\rangle\right] \tilde{e}_\nu\,, \qquad \text{with} \quad m_{\tilde{d}} := \tfrac{\tilde{d}^{m+1}-1}{\tilde{d}-1},$$

is as defined in (92), that is with the coefficient matrix $\big(\beta^{(\lambda_2)}_{j\nu}\big) =: \boldsymbol{\beta}_{\lambda_2}$ solving

$$(\boldsymbol{A}_{\mathfrak{Z}}^\top \boldsymbol{A}_{\mathfrak{Z}} + N\lambda_2 \boldsymbol{A}_{\mathfrak{Z}})\boldsymbol{\beta}_{\lambda_2} = \boldsymbol{A}_{\mathfrak{Z}} \boldsymbol{C}_{\mathfrak{Z}} \qquad \text{for} \quad \boldsymbol{C}_{\mathfrak{Z}} := \big(\big\langle \underline{\mathfrak{sig}}_\Lambda(Y^{(i)}), \eta^{-1}(j)\big\rangle\big)_{(i,j) \in [N] \times [m_{\tilde{d}}]}\,;$$

here, the map $\pi^{\mathrm{lex}}_{[m]} := \eta \circ \pi_{[m]} : \mathcal{H}_{\mathcal{Y}} \to \mathbb{R}^{m_{\tilde{d}}}$ is the shortlex-ordered projection onto $\mathcal{H}^{[m]}_{\mathcal{Y}}$.

The function defined in equation (107) serves as a data-based approximation of the (regression) function $\psi_f \in L^2(\mathbb{P}_X; \mathbb{R}^k)$ such that $\psi_f(X) = \mathbb{E}[f(Y)|X]$. Specifically, we anticipate that:

$$\mathbb{E}[f(Y) \,|\, X] \;\approx\; \hat{\mathscr{R}}^{\mathrm{I}}_{\underline{\lambda},f}[n, N, m](X), \tag{108}$$

where this approximation is expected to improve with sufficiently large values of n, N and with sufficiently small values of λ_1, λ_2. The precise quality of the approximation (108), and the requisite parameter choices for achieving this approximation within given error bounds, are established in Theorem 5.9 below, based on the threshold quantities defined in equations (104), (105), and (106).

In summary, the above combines to the following estimator for the regression operator (118).

Algorithm I: Computing the Regression Operator $f \mapsto \mu_\cdot(f)$ via Signatures

Goal: For processes X in $\mathbb{R}^{d_\mathcal{X}}$ and Y in $\mathbb{R}^{d_\mathcal{Y}}$, compute $\mathcal{L}^2(\mathbb{P}_Y;\mathbb{R}^k) \ni f \mapsto \psi_f \in \mathcal{L}^2(\mathbb{P}_X;\mathbb{R}^k)$, for regression functions $\psi_f := \mu_\cdot(f)$ (with $\mu := \mathbb{P}_{(X,Y)}$) such that $\psi_f(X) = \mathbb{E}[f(Y)\,|\,X]$.

1. **Input:** Samples $\mathfrak{z} := \{(x_j, y_j)\,|\,j \in [N]\}$ and $\mathfrak{y}_f := \big\{(y_l, f(y_l))\,|\,l \in [n]\big\}$ of the associations (X,Y) and $(Y, f(Y))$, for $f \equiv (f_1, \cdots, f_k) \in \mathcal{L}^2(\mathbb{P}_Y;\mathbb{R}^k)$, respectively.

   Hyperparameters: tensor normalisations (Ξ, Λ), resolution parameters $\lambda_f, \lambda_\Upsilon, m > 0$.

2. Compute the weights $\boldsymbol{\beta}_{\lambda_\Upsilon} \equiv \big(\beta_{j\nu}^{(\lambda_\Upsilon)}\big) \in \mathbb{R}^{N \times m_{d_\mathcal{Y}}}$ and $\boldsymbol{\alpha}^f_{\lambda_f} \equiv \big(\alpha_{li}^{(\lambda_f)}\big) \in \mathbb{R}^{n\times k}$ solving

$$\begin{aligned}(\boldsymbol{A}_\mathfrak{z}^\top \boldsymbol{A}_\mathfrak{z} + N\lambda_\Upsilon \boldsymbol{A}_\mathfrak{z})\boldsymbol{\beta}_{\lambda_\Upsilon} &= \boldsymbol{A}_\mathfrak{z}\boldsymbol{C}_\mathfrak{z},\\ (\boldsymbol{A}_\mathfrak{y}^\top \boldsymbol{A}_\mathfrak{y} + n\lambda_f \boldsymbol{A}_\mathfrak{y})\,\boldsymbol{\alpha}^f_{\lambda_f} &= \boldsymbol{A}_\mathfrak{y} B_{\mathfrak{y},f},\end{aligned} \tag{109}$$

   with $m_{d_\mathcal{Y}} := (\underline{d}_\mathcal{Y}^{m+1} - 1)/(\underline{d}_\mathcal{Y} - 1)$ and for the data-dependent coefficient matrices

$$\begin{aligned}\boldsymbol{A}_\mathfrak{z} &:= \Big(\big\langle \underline{\mathfrak{sig}}_\Xi(x_i), \underline{\mathfrak{sig}}_\Xi(x_j)\big\rangle\Big)_{i,j=1}^N \quad \text{and} \quad \boldsymbol{C}_\mathfrak{z} = \Big(\big\langle \underline{\mathfrak{sig}}_\Lambda(y_i), \mathrm{lex}^{-1}(j)\big\rangle\Big)_{(i,j)\in[N]\times[m_{d_\mathcal{Y}}]},\\ \boldsymbol{A}_\mathfrak{y} &:= \Big(\big\langle \underline{\mathfrak{sig}}_\Lambda(y_i), \underline{\mathfrak{sig}}_\Lambda(y_j)\big\rangle\Big)_{i,j=1}^n \quad \text{and} \quad \boldsymbol{B}_{\mathfrak{y},f} = \big(f_j(y_i)\big)_{(i,j)\in[n]\times[k]}.\end{aligned}$$

   Here, $\mathrm{lex} : [\underline{d}_\mathcal{Y}]^* \to \mathbb{N}_0$ is the shortlex ordering of the words in $[\underline{d}_\mathcal{Y}]^*$ (cf. footnote 7).

3. Compute the function, using the weights $\boldsymbol{\beta}_{\lambda_\Upsilon}$ and the standard basis $(\tilde{e}_\nu)$ of $\mathbb{R}^{m_{d_\mathcal{Y}}}$,

$$\hat{\Upsilon}^{\boldsymbol{\beta}_{\lambda_\Upsilon}}_{N,\lambda_\Upsilon,m} := \sum_{\nu=1}^{m_{d_\mathcal{Y}}} \left[\sum_{j=1}^N \beta_{j\nu}^{(\lambda_\Upsilon)} \langle \underline{\mathfrak{sig}}_\Xi(x_j), \underline{\mathfrak{sig}}_\Xi(\cdot)\rangle\right] \tilde{e}_\nu. \tag{110}$$

4. Compute the function, using the weights $\boldsymbol{\alpha}^f_{\lambda_f}$ and the quantity $\hat{\Upsilon}^{\boldsymbol{\beta}_{\lambda_\Upsilon}}_{N,\lambda_\Upsilon,m}$,

$$\hat{\psi}_f := \sum_{i=1}^k \left[\sum_{l=1}^n \alpha_{li}^{(\lambda_f)} \big\langle \pi^{\mathrm{lex}}_{[m]}(\underline{\mathfrak{sig}}_\Lambda(y_l)), \hat{\Upsilon}^{\boldsymbol{\beta}_{\lambda_\Upsilon}}_{N,\lambda_\Upsilon,m}(\cdot)\big\rangle_2\right] e_i \tag{111}$$

   for the standard basis (e_i) of $\mathbb{R}^k$ and with $\pi^{\mathrm{lex}}_{[m]} : \mathcal{H}_{d_\mathcal{Y}} \to \mathbb{R}^{m_{d_\mathcal{Y}}}$, $(q_w) \mapsto \sum_{\nu=1}^{m_{d_\mathcal{Y}}} q_{\mathrm{lex}^{-1}(\nu)} \tilde{e}_\nu$.

5. **Output:** $\hat{\psi}_f$ (estimator of the regression function ψ_f).

The fact that this estimator is a particular instance of Definition 2.6 is emphasised in Remark B.7.

Remark 5.7. The estimator (111), as well as the estimator (117) below, relies on the (bounded) signature kernels $\kappa_\Theta : \mathcal{Z} \times \mathcal{Z} \ni (z_1, z_2) \mapsto \langle \underline{\mathfrak{sig}}_\Theta(z_1), \underline{\mathfrak{sig}}_\Theta(z_2)\rangle \in \mathbb{R}$, where $(\Theta, \mathcal{Z}) \in \{(\Xi, \mathcal{X}), (\Lambda, \mathcal{Y})\}$, for both its parametrisation (109) and pointwise evaluation (110). These kernels, like the (bounded) signatures themselves [12, 26], can be efficiently computed or approximated with several specialised algorithms and estimators. For an overview of related works, see e.g. [8, 29, 47] and the references therein. ◆

5.4 Preliminary Results and Estimator for Finite-Dimensional Covariate

The scenario of Corollary 4.9, where we condition on X some finite-dimensional covariate $Z \in L^2(\mathbb{P};\mathbb{R}^k)$ instead of the Hilbert-valued covariate $\underline{\mathfrak{sig}}_\Lambda(Y)$, permits structural results similar to those in Lemma 5.5 and Propositions 5.1 and 5.3. These statements, given below, can be proven entirely in parallel to the aforementioned results and thus may be regarded as corollaries to the above proofs.

In a slight modification from the setting in (89), this Z-covaried scenario operates on the given data

$$\mathfrak{W} \equiv \big\{\big(X^{(j)}, (Z^1_{(j)}, \cdots, Z^k_{(j)})\big) \mid j \in [M]\big\} \qquad (M \in \mathbb{N}), \tag{112}$$

i.e., on some fixed ensemble of M-many iid copies of (X, Z). We can then observe the following.

Corollary 5.8. *Let X be as in (52), and consider the subspace $\mathfrak{h}_k \coloneqq \big\{h \equiv (h^1, \cdots, h^k) : \mathcal{X} \to \mathbb{R}^k \,\big|\, h^i \in \mathcal{H}_\kappa,\ \forall\, i \in [k]\big\} \subset L^2(\mathbb{P}_X;\mathbb{R}^k)$ together with the objectives*

$$\begin{aligned}
\boldsymbol{\phi}_Z &: L^2\big(\mathbb{P}_X;\mathbb{R}^k\big) \longrightarrow \mathbb{R}_+, \quad && g \mapsto \int_{\mathcal{X}\times\mathbb{R}^k} \big|z - g(x)\big|^2\, \mathbb{P}_{(X,Z)}(\mathrm{d}x, \mathrm{d}z),\\
\boldsymbol{\phi}_{Z|\lambda} &: \mathfrak{h}_k \longrightarrow \mathbb{R}_+, && g \mapsto \boldsymbol{\phi}_Z(g) + \lambda\|g\|^2_{\mathfrak{h}_k},\\
\hat{\boldsymbol{\phi}}_{\mathfrak{W},\lambda} &: \mathfrak{h}_k \longrightarrow \mathbb{R}_+, && g \mapsto \int_{\mathcal{X}\times\mathbb{R}^k} \big|z - g(x)\big|^2\, \hat{\mu}_{\mathfrak{W}}(\mathrm{d}x, \mathrm{d}z) + \lambda\|g\|^2_{\mathfrak{h}_k},
\end{aligned}$$

for $\lambda > 0$ and where $\hat{\mu}_{\mathfrak{W}} \coloneqq \frac{1}{M}\sum_{j\in[M]} \delta_{(X^{(j)}, Z^{(j)})}$. Then both the minimizer $\Upsilon^*_{Z,\lambda}$ of $\boldsymbol{\phi}_{Z|\lambda}$ and

$$\hat{\Upsilon}^*_{Z,\lambda;M} \coloneqq \operatorname*{arg\,min}_{h\in\mathfrak{h}_k} \hat{\boldsymbol{\phi}}_{\mathfrak{W},\lambda}(h) \quad \textit{are} \quad \textit{unique},$$

and the latter minimizer admits the representation (with $(e_i \mid i \in [k])$ the standard basis of $\mathbb{R}^k$)

$$\hat{\Upsilon}^*_{Z,\lambda;M} = \sum_{i=1}^{k}\left[\sum_{j=1}^{M} \hat{\alpha}^{(\lambda)}_{ji} \big\langle \underline{\mathfrak{sig}}_\Xi(X^{(j)}), \underline{\mathfrak{sig}}_\Xi(\cdot)\big\rangle\right] e_i \tag{113}$$

for any coefficient matrix $\hat{A}_\star \equiv \big(\hat{\alpha}^{(\lambda)}_{ji}\big) \in \mathbb{R}^{M\times k}$ which, abbreviating $\tilde{q}_j \coloneqq \underline{\mathfrak{sig}}_\Xi(X^{(j)})$, solves

$$(\boldsymbol{A}^\top_{\mathfrak{W}}\boldsymbol{A}_{\mathfrak{W}} + M\lambda\boldsymbol{A}_{\mathfrak{W}})\hat{A}_* = \boldsymbol{A}_{\mathfrak{W}}\boldsymbol{B}_{\mathfrak{W}} \quad \textit{for} \quad \boldsymbol{A}_{\mathfrak{W}} \coloneqq (\langle \tilde{q}_i, \tilde{q}_j\rangle)_{i,j=1}^M \textit{ and } \boldsymbol{B}_{\mathfrak{W}} \coloneqq \big(Z^j_{(i)}\big)_{(i,j)\in[M]\times[k]}. \tag{114}$$

Then the following holds:

(i) *the space $\mathfrak{h}_k$ is a separable vRKHS with reproducing kernel $\mathcal{K} : \mathcal{X}^{\times 2} \to \mathfrak{L}(\mathbb{R}^k)$, $(x, y) \mapsto \kappa(x, y)\mathrm{Id}_{k\times k}$, and also a $\|\cdot\|_{L^2(\mathbb{P}_X;\mathbb{R}^k)}$-dense subspace of $L^2(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P}_X;\mathbb{R}^k)$;*

(ii) *for (the unique) $\psi_{X,Z} \in L^2(\mathbb{P}_X;\mathbb{R}^k)$ such that $\mathbb{E}[Z|X] = \psi_{X,Z}(X)$ (Doob-Dynkin) and the modulus[9] $\hat{\vartheta}_{\psi_{X,Z}}(\tilde{\varepsilon}) \coloneqq \frac{\tilde{\varepsilon}^2}{4}\big(\inf\big\{\|h\|^2_{\mathfrak{h}_k} \,\big|\, h \in \mathfrak{h}_k : \big\|\psi_{X,Z} - h\big\|^2_{L^2(\mathbb{P}_X;\mathbb{R}^k)} \le \tilde{\varepsilon}^2/2\big\}\big)^{-1}$, we have that*

$$\sup\big\{\rho^{\mathrm{I}}_Z(\lambda) \coloneqq \big\|\psi_{X,Z} - \Upsilon^*_{Z,\lambda}\big\|_{L^2(\mathbb{P}_X;\mathbb{R}^k)} \,\big|\, 0 < \lambda \le^{10} \hat{\vartheta}_{\psi_{X,Z}}(\epsilon)\big\} \le \epsilon, \quad \textit{for all} \ \ \epsilon > 0\,;$$

(iii) *for any $\delta > 0$, there is an explicit monotone null sequence $(\tilde{\beta}_M) \equiv (\tilde{\beta}_M(\delta, \lambda))_{M\in\mathbb{N}} \subset \mathbb{R}_+$, depending on (Ξ, λ, δ), such that for each $M \in \mathbb{N}$ we have*

$$\mathbb{P}\big(\rho^{\mathrm{II}}_{Z,\lambda}(M) \ge \tilde{\beta}_M\big) \le \delta \qquad \textit{for} \quad \rho^{\mathrm{II}}_{Z,\lambda}(M) \coloneqq \big\|\Upsilon^*_{Z,\lambda} - \hat{\Upsilon}^*_{Z,\lambda;M}\big\|_{L^2(\mathbb{P}_X;\mathbb{R}^k)}.$$

[9] We set $\hat{\vartheta}_{\psi_{X,Z}}(\tilde{\epsilon}) \coloneqq \infty$ if the defining infimum is zero. [10] If $\hat{\vartheta}_{\psi_{X,Z}}(\epsilon) = \infty$, then the constraint '$\lambda \le \hat{\vartheta}_{\psi_{X,Z}}(\epsilon)$' is to be dropped.

Proof. The assertions concerning $\mathfrak{h}_k$ are derived entirely analogously to the corresponding properties of (80), as demonstrated in the proof of Proposition 5.1. The uniqueness of $(\hat{\Upsilon}_{Z,\lambda;M})$ is established analogously to the uniqueness of $\Upsilon^*_{\mu,\lambda,m}$ in (86), while its data-based representation in (113) is derived similarly to the $\hat{\Upsilon}^{\mathrm{lex}}_{N,\lambda,m}$-representation in (92). Statement (ii) is derived in the same manner as Lemma 5.5 (iii), while the corollary's statement (iii) holds, for example, for the sequence $(\tilde{\beta}_M) \coloneqq \big(\frac{\sqrt{c_\Xi}\|Z\|_{L^2(\mathbb{P};\mathbb{R}^k)}}{\lambda\sqrt{M\delta}}\big)$, as established by the same reasoning as in the proof of Lemma 5.5 (v). □

Similar to how the results of Section 5.2 informed the construction of the estimator (107), setting

$$\hat{\mathscr{R}}^{\mathrm{II}}_{\lambda,Z}[M] \coloneqq \sum_{i=1}^{k}\left[\sum_{j=1}^{M}\hat{\alpha}^{(\lambda)}_{ij}\langle \underline{\mathfrak{sig}}_\Xi(X^{(j)}), \underline{\mathfrak{sig}}_\Xi(\cdot)\rangle\right] e_i \,:\, \mathcal{X} \to \mathbb{R}^k \tag{115}$$

yields an estimator of the regression function between Z and X. Through the data (112), this estimator relies on a parameter $\lambda > 0$ and coefficients $\hat{A}_\star = \big(\hat{\alpha}^{(\lambda)}_{ji}\big) \in \mathbb{R}^{M\times k}$ solving (114).

When applied to the vector $Z \coloneqq f(Y)$, the function (115) is designed to yield an approximation

$$\mathbb{E}[f(Y)\,|\,X] \;\approx\; \hat{\mathscr{R}}^{\mathrm{II}}_{\lambda,Z}[M](X), \tag{116}$$

similar to (108), where this approximation is expected to improve with sufficiently large M and with sufficiently small λ. The precise quality of the approximation (116), and the requisite parameter choices for achieving this approximation within given error bounds, are established in Theorem 5.9 below, based on the threshold quantities defined in Corollary 5.8.

The results of Section 5 combine to the following estimators for the regression operator (118).

Algorithm II: Computing $\psi_Z \in \mathcal{L}^2(\mathbb{P}_X;\mathbb{R}^k)$ such that $\psi_Z(X) = \mathbb{E}[Z\,|\,X]$ $(Z \in L^2(\mathbb{P};\mathbb{R}^k))$.

Goal: For a stochastic process X in $\mathbb{R}^d$ and a random vector Z in $\mathbb{R}^k$, estimate a regression function $\psi_Z \in \mathcal{L}^2(\mathbb{P}_X;\mathbb{R}^k)$ such that $\psi_Z(X) = \mathbb{E}[Z\,|\,X]$.

1. **Input:** Samples $\mathfrak{w} \coloneqq \big\{\big(x_i, (z_i^1,\dots,z_i^k)\big)\,\big|\,i\in[M]\big\}$ of the association (X,Z).

   Hyperparameters: tensor normalisation $\Xi : \mathcal{H}^{\downarrow}_{\underline{d}} \to \mathcal{H}_{\underline{d}}$, regularity parameter $\lambda > 0$.

2. Compute the weights $\hat{\boldsymbol{\alpha}}_\lambda \equiv \big(\hat{\alpha}^{(\lambda)}_{ij}\big) \in \mathbb{R}^{k\times M}$ by solving

   $$(\boldsymbol{A}_\mathfrak{w}^\top \boldsymbol{A}_\mathfrak{w} + M\lambda \boldsymbol{A}_\mathfrak{w})\hat{\boldsymbol{\alpha}}_\lambda = \boldsymbol{A}_\mathfrak{w}\boldsymbol{B}_\mathfrak{w}$$

   for the data-dependent coefficient matrices

   $$\boldsymbol{A}_\mathfrak{w} \coloneqq \Big(\langle \underline{\mathfrak{sig}}_\Xi(x_i), \underline{\mathfrak{sig}}_\Xi(x_j)\rangle\Big)_{i,j=1}^{M} \quad\text{and}\quad \boldsymbol{B}_\mathfrak{w} = \big(z_i^j\big)_{(i,j)\in[M]\times[k]}.$$

3. Compute the function, using the weights $\hat{\boldsymbol{\alpha}}_\lambda$ and the standard basis (e_i) of $\mathbb{R}^k$,

   $$\hat{\psi}_Z \;\coloneqq\; \sum_{i=1}^{k}\left[\sum_{j=1}^{M}\hat{\alpha}^{(\lambda)}_{ij}\langle \underline{\mathfrak{sig}}_\Xi(x_j), \underline{\mathfrak{sig}}_\Xi(\cdot)\rangle\right] e_i. \tag{117}$$

4. **Output:** $\hat{\psi}_Z$ (estimator of the regression function ψ_Z).

The fact that this estimator is a particular instance of Definition 2.6 is emphasised in Remark B.8.

5.5 Convex and Consistent Stochastic Process Regression

The main objective of this paper is to establish a universally consistent framework for the nonparametric regression of stochastic processes. In Section 2.3, particularly through Proposition 2.4 and Definition 2.6, we have have identified as our central target object for this the regression operator

$$\mathfrak{c}_\xi(\mu) \,:\, L^2(\mu_{\mathcal{Y}}) \longrightarrow L^2(\xi), \quad f \longmapsto \mu_{\cdot}(f), \tag{118}$$

defined for any measure $\mu \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ with $\mathcal{X}$-marginal $\xi := \mu_{\mathcal{X}} \in \mathcal{M}_1(\mathcal{X})$. Recall from (20) and (21) that if μ is the joint law of two stochastic processes X and Y as in (52), i.e., $\mu = \mathbb{P}_{(X,Y)}$ and thus $\xi = \mathbb{P}_X$, then (118) characterises the conditional statistical dependence of Y given X.

Assuming $\mu = \mathbb{P}_{(X,Y)}$ entails no loss of generality (Remark 2.3), and the samples of μ then read:

$$\mathfrak{Z} \equiv \big\{(X^{(j)}, Y^{(j)}) \;\big|\; j \in [N]\big\} \qquad (N \in \mathbb{N}) \tag{119}$$

as specified in (89) (cf. (10)), or they are of the form (112) if we are interested in dependencies modelled by $\mu = \mathbb{P}_{(X,Z)}$ for a random covariate Z in $\mathbb{R}^k$. To enable parallel learning of the arguments f in (118), we may also wish to observe some marginal samples of $\mu_{\mathcal{Y}}$, which appear as $\mathfrak{Y}$ in (89).

We then have the following result on the approximation quality of the structures (107) and (115).

Theorem 5.9. *Consider the estimators $\hat{\mathscr{R}}^\nu_\theta : \mathcal{L}^2(\mathbb{P}_Y; \mathbb{R}^k) \to \mathcal{L}^2(\mathbb{P}_X; \mathbb{R}^k)$ $(\nu = \mathrm{I}, \mathrm{II})$ given by*

$$\hat{\mathscr{R}}^{\mathrm{I}}_{(\lambda,n,N,m)}[f] := \hat{\mathscr{R}}^{\mathrm{I}}_{\lambda,f}[n,N,m] \qquad \text{and} \qquad \hat{\mathscr{R}}^{\mathrm{II}}_{(\lambda,M)}[f] := \hat{\mathscr{R}}^{\mathrm{II}}_{\lambda,f(Y)}[M], \tag{120}$$

as defined in (107) and (115). Then for each $f \in \mathcal{L}^2(\mathbb{P}_Y; \mathbb{R}^k)$, any accuracy $\varepsilon > 0$ and any $\delta > 0$,

$$\sup_{\theta \,\in\, \mathcal{A}^\nu_f(\varepsilon,\delta)} \mathbb{P}\Big(\big\|\mu_{\cdot}(f) - \hat{\mathscr{R}}^\nu_\theta[f]\big\|_{L^2(\mathbb{P}_X;\mathbb{R}^k)} \geq \varepsilon\Big) \;\leq\; \delta \tag{121}$$

for both $\nu = \mathrm{I}, \mathrm{II}$. In particular, for each $\varepsilon > 0$ and any given confidence level $q \in [0,1)$, we have

$$\inf_{\theta \,\in\, \Theta^\nu_f(\varepsilon,q)} \mathbb{P}\Big(\big|\mathbb{E}\big[f(Y)\,\big|\,X\big] - \hat{\mathscr{R}}^\nu_\theta[f](X)\big| < \varepsilon\Big) \;\geq\; q \tag{122}$$

for both $\nu = \mathrm{I}, \mathrm{II}$. The inequalities (121) hold for the parameter regimes

$$\begin{aligned}\mathcal{A}^{\mathrm{I}}_f(\varepsilon,\delta) := \big\{\big((\lambda_1,\lambda_2),n,N,m\big) \in \mathbb{R}^2_{>0} \times \mathbb{N}^3 \;\big|\; &\lambda_1 \leq \lambda^{\mathrm{I}}_f(\varepsilon),\; n \geq n_{f,\lambda_1}(\varepsilon,\delta),\; m \geq m_{f,\lambda_1}(\varepsilon;\mathfrak{Y}),\\ &\lambda_2 \leq \lambda^{\mathrm{II}}_{\lambda_1,m}(\varepsilon;\mathfrak{Y}),\; N \geq N_{\lambda_1,\lambda_2,m}(\varepsilon,\delta;\mathfrak{Y})\big\}\end{aligned}$$

$$\text{and} \qquad \mathcal{A}^{\mathrm{II}}_f(\varepsilon,\delta) := \big\{(\lambda,M) \in \mathbb{R}_{>0} \times \mathbb{N} \;\big|\; \lambda \leq \hat{\vartheta}_{\psi_{X,Z}}(\varepsilon/2) \;\text{ and }\; \tilde{\beta}_M(\delta,\lambda) \leq \varepsilon/2\big\},$$

respectively, defined for any $(\tilde{\beta}_m)$ as in Cor. 5.8 (iii) *and in the notation of (104), (105), (106). The inequalities (122) hold for the parameter regimes*

$$\Theta^{\mathrm{I}}_f(\varepsilon,q) := \bigcup_{(\tilde{\varepsilon},\tilde{\delta}) \in I(\varepsilon,q)} \mathcal{A}^{\mathrm{I}}_f(\tilde{\varepsilon},\tilde{\delta}) \qquad \text{and} \qquad \Theta^{\mathrm{II}}_f(\varepsilon,q) := \bigcup_{(\tilde{\varepsilon},\tilde{\delta}) \in I(\varepsilon,q)} \mathcal{A}^{\mathrm{II}}_f(\tilde{\varepsilon},\tilde{\delta}),$$

respectively, which are defined over $I(\varepsilon,q) := \{(\tilde{\varepsilon},\tilde{\delta}) \in \mathbb{R}^2_{>0} \mid \tilde{\varepsilon}(1-\tilde{\delta}) + \tilde{\delta}\varepsilon^2 \;\leq\; (1-q)\varepsilon^2\}$.

Proof. The consistency guarantees (121) and (122) rely on a combination of the results from Sections 5.1 to 5.4. A detailed proof of Theorem 5.9 is provided in Appendix B.3.6. □

Remark 5.10 (Time-Discretized Non-iid Data)**.** For simplicity, Theorem 5.9 is based on the standard assumption that the empirical data (119) and (112), from which the estimators (120) are computed, are iid samples of the process pairs (X, Y) and (X, Z), respectively. Yet in practice one usually has access to time-discretized samples of these processes only, often taken over non-equally spaced time grids, and oftentimes only statistically dependent (time-discretized) sample paths of these pairs are available rather than many independent realisations. The above consistency theorem can be extended to hold under these non-standard assumptions, in line with the approaches established in [49, Section 8]. A full elaboration of according modifications is left for future research in order to not exceed the scope of this paper. ◆

6 Numerical Experiments

Numerical experiments and practical example applications will be included in the forthcoming arXiv version of this paper.

Appendix A Notation

Below are some of the symbols and terminology that we use throughout the main paper.

Symbol	Meaning	Page
$\mathcal{X}, \mathcal{Y}$	factor spaces of the relation space $\mathcal{X} \times \mathcal{Y}$, usually assumed to be Polish spaces.	4
$\mathcal{B}(\mathcal{Z})$	Borel σ-algebra of a given topological space $\mathcal{Z}$.	5
$\mathcal{M}_1(\mathcal{Z})$	$\coloneqq \{\upsilon : \mathcal{B}(\mathcal{Z}) \to [0,1] \,\vert\, \upsilon \text{ measure}\}$; cone of all Borel probability measures on $\mathcal{Z}$.	5
$\mathfrak{c}_\xi(\mu)$	the conditional dependence of a measure $\mu \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ wrt. the source law $\xi \in \mathcal{M}_1(\mathcal{X})$; often identified with the regression operator (14), denoted by the same symbol.	6
$\mu_{\cdot}(f)$	$: \mathcal{X} \ni x \mapsto \int_{\mathcal{Y}} f \,\mathrm{d}\mu_x$; (conditional) expectation of $f \in \mathcal{L}^2(\xi)$ wrt. the ξ-disintegration $\mu_{\cdot} = (\mu_x)_{x \in \mathcal{X}}$ of a measure $\mu \in \mathcal{M}_1^\xi(\mathcal{X} \times \mathcal{Y})$.	6
δ_p	$: 2^{\mathcal{Z}} \ni A \mapsto \mathbb{1}_A(p) \in [0,1]$; the Dirac measure at a given point $p \in \mathcal{Z}$.	6
$\mathcal{L}^p(\upsilon; E)$	$\coloneqq \{f : \mathcal{Z} \to E \,\vert\, f \text{ is measurable, } \int_{\mathcal{Z}} \Vert f \Vert_E^p \,\mathrm{d}\upsilon < \infty\}$, $p \in [1, \infty) \cup \{0\}$.	7
$\mathbb{1}_A$	$z \mapsto \mathbb{1}_A(z) \coloneqq \{1 \text{ if } z \in A; 0 \text{ else}$; indicator function over a set A.	8
$\mathcal{C}_d$	$\coloneqq \{x : [0,1] \to \mathbb{R}^d \mid x \text{ is continuous}\}$; space of continuous paths in $\mathbb{R}^d$, usually endowed with the uniform norm $\Vert x \Vert_\infty \coloneqq \sup_{t \in [0,1]} \vert x_t \vert$ (unless otherwise stated).	10
$\mathcal{C}_d^1$	the space of all absolutely continuous paths in $\mathbb{R}^d$.	10
$\vert \cdot \vert$	$: \mathbb{R}^n \to \mathbb{R}_+$, Euclidean norm on $\mathbb{R}^n$; in partic., $\vert \cdot \vert$ is the absolute value on $\mathbb{R} = \mathbb{R}^1$.	10
$\alpha \wedge \beta$	$\coloneqq \min\{\alpha, \beta\}$; the minimum of $\alpha, \beta \in \mathbb{R}$.	12
$\delta_{a,b}$	$\coloneqq \{1 \text{ if } a = b; 0 \text{ else}$; the Kronecker delta over two elements a, b.	12
$[k]$	$\coloneqq \{1, \ldots, k\}$, and $[k]_0 \coloneqq [k] \cup \{0\}$ $(k \in \mathbb{N})$.	12
$[d]^\star$	$\coloneqq \bigcup_{m \geq 0} [d]^{\times m} = \{\emptyset, 1, 2, \ldots, 11, 12, 112, \ldots\}$; the set of all multi-indices with entries in $[d]$, where $[d]^0 \coloneqq \{\emptyset\}$ and $i_1 \cdots i_m \equiv (i_1, \ldots, i_m)$.	13
$\underline{\mathtt{sig}}$	the time-augmented signature transform.	14
$\mathcal{H}_d$	codomain of the signature on d-dimensional paths, see (42) for the definition.	15
$\overline{\mathbb{R}}$	$\coloneqq \mathbb{R} \cup \{-\infty, \infty\}$; the affine closure of the real numbers.	14
$\underline{\mathtt{sig}}_\Lambda, \underline{\mathtt{sig}}_\Xi$	bounded signature transforms, for feature normalisations Λ and Ξ.	16
$L_X^p(\mathcal{H})$	space of all p-integrable, $\mathcal{H}$-valued X-measurable functions, see (57).	19
$(e_i)_{i \in [k]}$	the standard basis of $\mathbb{R}^k$ (i.e.: $e_i = (\delta_{1i}, \cdots, \delta_{ki}) \in \mathbb{R}^k$, for each $i \in [k]$).	21
$\mathrm{supp}(\upsilon)$	the support of a measure $\upsilon \in \mathcal{M}_1(\mathcal{Z})$, i.e. the smallest closed subset $C \subseteq \mathcal{Z}$ with $\upsilon(C) = 1$ (see e.g. [23, Lemma 1.19]).	23
$\varphi\big\vert_{\tilde{A}}$	the restriction of a map $\varphi : A \to B$ to a subdomain $\tilde{A} \subseteq A$.	24
$\langle u, v \rangle_2$	$\coloneqq u_1 v_1 + \ldots + u_d v_d$; the dot product of two vectors $u = (u_i), v = (v_i)$ in $\mathbb{R}^d$.	51

Notation A.1. Unless specified otherwise, we assume that both $\mathcal{X}$ and $\mathcal{Y}$ are Polish spaces. For a given topological space $\mathcal{Z}$, we denote by $\mathcal{B}(\mathcal{Z})$ its Borel σ-algebra and by $\mathcal{M}_1(\mathcal{Z}) \coloneqq \{\upsilon : \mathcal{B}(\mathcal{Z}) \to [0,1] \,|\, \upsilon \text{ measure}\}$ the cone of all Borel probability measures on $\mathcal{Z}$. Given a measure $\upsilon \in \mathcal{M}_1(\mathcal{Z})$, we write $\mathrm{supp}(\upsilon)$ for its support, that is the smallest closed subset $C \subseteq \mathcal{Z}$ with $\upsilon(C) = 1$ (e.g. [23, Lemma 1.19]). The Cartesian product $\mathcal{X} \times \mathcal{Y}$ is endowed with the product topology (thus, $\mathcal{B}(\mathcal{X} \times \mathcal{Y}) = \mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{Y})$), the canonical coordinate projections $\mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ and $\mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ are denoted by $\hat{\pi}_1$ and $\hat{\pi}_2$ respectively, and for a given $\mu \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ we write $\mu_{\mathcal{X}} \coloneqq (\hat{\pi}_1)_* \mu \equiv \mu \circ \hat{\pi}_1^{-1}$ and $\mu_{\mathcal{Y}} \coloneqq (\hat{\pi}_2)_* \mu$ for its marginals. Given $\xi \in \mathcal{M}_1(\mathcal{X})$, we write $\mathcal{M}_1^{\xi}(\mathcal{X} \times \mathcal{Y}) \coloneqq \{\tilde{\mu} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y}) \mid \tilde{\mu}_{\mathcal{X}} = \xi\}$ for the set of all measures on $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{X}$-marginal ξ. The law of some $\mathcal{Z}$-valued (Borel) random variable Z over a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ is denoted by $\mathbb{P}_Z \coloneqq Z_* \mathbb{P} \equiv \mathbb{P} \circ Z^{-1} \in \mathcal{M}_1(\mathcal{Z})$. ◆

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] C. Carmeli, E. De Vito, A. Toigo, and V. Umanitá. Vector Valued Reproducing Kernel Hilbert Spaces and Universality. *Analysis and Applications*, **8**.01:19–61, 2010.

[8] T. Cass, T. Lyons, and X. Xu. Weighted Signature Kernels. *Ann. Appl. Probab.*, 34(1A):585–626, 2024.

[9] 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.

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

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

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

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

[14] D. L. Cohn. *Measure Theory*. Second Edition. Birkhäuser Advanced Texts, 2013.

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

[16] F. Cucker and S. Smale. On the Mathematical Foundations of Learning. *Bull. Amer. Math. Soc.*, **39**.1:1–49, 2002.

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

[18] 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.

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

[20] 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.

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

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

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

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

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

[26] P. Kidger and T. Lyons. Signatory: differentiable computations of the signature and logsignature transforms, on both CPU and GPU. *arXiv: 2001. 00706 and https: // github. com/ patrick-kidger/ signatory ; published at ICLR 2021*, 2020.

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

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

[29] D. Lee and H. Oberhauser. The Signature Kernel. *arXiv preprint arXiv:2305.04625*, 2023.

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

[31] 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.

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

[33] Z. Li, D. Meunier, M. Mollenhauer, and A. Gretton. Optimal Rates for Regularized Conditional Mean Embedding Learning. *Advances in Neural Information Processing Systems*, 35:4433–4445, 2022.

[34] 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.

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

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

[37] 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.

[38] 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.

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

[40] H. Owhadi and C. Scovel. Separability of Reproducing Kernel Spaces. *Proc. Amer. Math. Soc.*, **145**.5:2131–2138, 2017.

[41] J. Park and K. Muandet. A Measure-Theoretic Approach to Kernel Conditional Mean Embeddings. *Advances in neural information processing systems*, 33:21247–21259, 2020.

[42] J. Park and K. Muandet. Regularised Least-Squares Regression With Infinite-Dimensional Output Space. *arXiv preprint arXiv:2010.10973*, 2020.

[43] R. Penrose. On Best Approximate Solutions of Linear Matrix Equations. In *Math. Proc. Cambridge Philos. Soc.*, volume **52**.1, pages 17–19. Cambridge University Press, 1956.

[44] 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.

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

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

[47] C. Salvi, T. Cass, J. Foster, T. Lyons, and W. Yang. The Signature Kernel is the Solution of a Goursat PDE. *SIAM Journal on Mathematics of Data Science*, **3**.3:873–899, 2021.

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

[49] A. Schell and H. Oberhauser. Nonlinear Independent Component Analysis for Discrete-Time and Continuous-Time Signals. (Reference is made to arXiv:2102.02876). *Ann. Statist.*, **51**.2:487–518, 2023.

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

[51] B. Schölkopf, R. Herbrich, and A. J. Smola. A Generalized Representer Theorem. In *Computational Learning Theory*, pages 416–426. Lecture Notes in Computer Science. Vol. **2111**. Springer, 2001.

[52] L. Schwartz. Sous-Espaces Hilbertiens d'Espaces Vectoriels Topologiques et Noyaux Associés (Noyaux Reproduisants). *J. Analyse Math.*, **13**:115–256, 1964.

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

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

[55] 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.

[56] M. Sipser. *Introduction to the Theory of Computation*. (3 ed.). Boston, MA: Cengage Learning, 2013.

[57] T.O. To and K.W. Yip. A generalized Jensen's inequality. *Pacific J. Math.*, **58**.1:255–259, 1975.

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

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

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

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

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

Appendix B Further Proofs and Technical Remarks

B.1 Ad Section 2

B.1.1 Remarks on Product-Space Measures and Operator Parametrizations

Remark B.1. For each (Borel) probability measure $\mu \in \mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$ with $\mathcal{X}$-marginal $\xi \in \mathcal{M}_1(\mathcal{X})$, there is a canonical pair (X,Y) of jointly distributed (Borel) rvs $X:\Omega\to\mathcal{X}$ and $Y:\Omega\to\mathcal{Y}$ such that $\mu = \mathbb{P}_{(X,Y)}$ and $\xi = \mathbb{P}_X$. Conversely, for any pair $(\tilde{X},\tilde{Y}):\Omega\to\mathcal{X}\times\mathcal{Y}$ of (jointly distributed Borel) random variables, their law $\mathbb{P}_{(\tilde{X},\tilde{Y})}$ is a Borel probability measure with $\mathcal{X}$-marginal $\mathbb{P}_{\tilde{X}}$.

Proof. This is trivial: Given $\mu\in\mathcal{M}_1^{\xi}(\mathcal{X}\times\mathcal{Y})$, set $(\tilde{\Omega},\tilde{\mathscr{F}},\tilde{\mathbb{P}}) \coloneqq (\mathcal{X}\times\mathcal{Y},\mathcal{B}(\mathcal{X}\times\mathcal{Y}),\mathfrak{c}_\xi(\mu)\otimes\xi)$, which defines a probability space, and consider the canonical projections $X\coloneqq\hat{\pi}_1$ and $Y\coloneqq\hat{\pi}_2$ (thus, $Z\coloneqq(X,Y) = \mathrm{id}_{\mathcal{X}}\times\mathrm{id}_{\mathcal{Y}}$). (Recall that $\hat{\pi}_1:\tilde{\Omega}\ni(x,y)\mapsto x\in\mathcal{X}$ and $\hat{\pi}_2:\tilde{\Omega}\ni(x,y)\mapsto y\in\mathcal{X}$.) Then $X:\tilde{\Omega}\to\mathcal{X}$ and $Y:\tilde{\Omega}\to\mathcal{Y}$ are (Borel) random variables with $\mathbb{P}_X=\xi$ and joint law $\mathbb{P}_Z=\mathbb{P}=\mu$. The converse statement is immediate from the definitions. ◆

Remark B.2 (Injective Parametrization (17) over a Hilbert Space)**.** Recognizing the importance of the injectivity of the parametrization in (semi-)parametric statistical models for the uniqueness, identifiability, and analytical tractability of a model, let us remark that the parametrization (17) of the conditional dependency $\mathfrak{c}_\xi(\mu)$ is 'injective' in the following basic sense: Denoting by $(\mathcal{H}_\kappa,\|\cdot\|_\kappa)\subset\mathcal{L}^2(\mu_{\mathcal{Y}})$ the (q-induced) RKHS with kernel $\kappa:(x,y)\mapsto\langle q(x),q(y)\rangle$, there is a 'parameter' Hilbert space $\tilde{\mathcal{H}}_q$ (that is conveniently explicit and regular for the maps q introduced in Sect. 3.3) such that

$$\mathfrak{c}_\xi(\mu)\ :\ \overline{\mathcal{H}_\kappa}^{L^2}\longrightarrow L^2(\xi)\quad\text{reads}\quad \mathfrak{c}_\xi(\mu)|_{\mathcal{H}_\kappa}=\psi_q\circ\iota \tag{123}$$

for an isometric isomorphism $\iota : \mathcal{H}_\kappa \to \tilde{\mathcal{H}}_q$ and some q-defined bounded linear operator $\psi_q : \tilde{\mathcal{H}}_q \to L^2(\xi)$, the '$q$-parametrisation' of $\mathfrak{c}_\xi(\mu)$, where both ι and ψ_q are canonically related to q.

Indeed: Take any $f \in \mathcal{H}_\kappa$, whence $f = \langle \ell_f, q(\cdot)\rangle$ for some $\ell_f \in \mathcal{H}_\mathcal{Y}$. Writing $[\ell_f] \coloneqq \big\{\ell \in \mathcal{H}_\mathcal{Y} \,\big|\, f = \langle \ell, q(\cdot)\rangle\big\}$, note that $[\ell_f] = \{\ell \in \mathcal{H}_\mathcal{Y} \mid \ell \sim \ell_f\} = \ell_f + F$ under the equivalence relation: $\ell \sim \tilde{\ell} :\Longleftrightarrow \langle \ell, q\rangle = \langle \tilde{\ell}, q\rangle \Longleftrightarrow (\ell - \tilde{\ell}) \in F \coloneqq \{\hat{\ell} \in \mathcal{H}_\mathcal{Y} \mid \langle \hat{\ell}, q(\cdot)\rangle = 0\} = \ker(\varphi)$, for the bounded linear surjection $\varphi : \mathcal{H}_\mathcal{Y} \twoheadrightarrow \mathcal{H}_\kappa$, $\hat{\ell} \mapsto \langle \hat{\ell}, q(\cdot)\rangle$. Consequently, $\iota : \big(\mathcal{H}_\kappa, \|\cdot\|_\kappa\big) \ni f \mapsto [\ell_f] \in \tilde{\mathcal{H}}_q \coloneqq (\mathcal{H}_\mathcal{Y}/F, |||\cdot|||)$, with $|||[\ell]||| \coloneqq \inf\{\|\tilde{\ell}\|_{\mathcal{H}_\mathcal{Y}} \mid \tilde{\ell} \in [\ell]\}$ the canonical quotient space norm, is a (norm)isometric isomorphism of Hilbert spaces: injectivity is clear, while the definition of the RKHS-norm $\|\cdot\|_\kappa$ yields norm-preservation, which in turn (shows that $|||\cdot|||$ preserves the parallelogram identity and hence) reveals $|||\cdot|||$ as an inner product norm; that the quotient $\tilde{H}$ is complete follows since $F\,(= \ker(\varphi))$ is closed in $\mathcal{H}_\mathcal{Y}$. Now the (linear, and bounded due to (16)) evaluation map $\psi_q : \tilde{\mathcal{H}}_q \ni [\ell_f] \mapsto \langle \ell_f, q(\cdot)\rangle \in L^2(\xi)$ is well-defined since clearly $\psi_q(\ell_f) = \psi_q(\ell)$ for each $\ell \in [\ell_f]$, and so the asserted composition (123) holds with Proposition 2.4 (17) and said proposition's final statement. ◆

B.1.2 Proof of Proposition 2.4

Proof. First, it is clear that each measure $\mu_x \in \mathcal{M}_1(\mathcal{Y})$ can be identified with the linear functional $\mu_x : L^2(\mu_\mathcal{Y}) \ni f \mapsto \int_\mathcal{Y} f \,\mathrm{d}\mu_x \in \overline{\mathbb{R}}$ (trivially, since $\{\mathbb{1}_A \mid A \in \mathcal{B}(\mathcal{Y})\} \subset L^2(\mu_\mathcal{Y})$). That the functions $\mu_\cdot(f) : \mathcal{X} \ni x \mapsto \mu_x(f)$ are in $L^2(\xi)$, for each $f \in L^2(\mu_\mathcal{Y})$, follows from Jensen's inequality:

$$\big\|\mu_\cdot(f)\big\|^2_{L^2(\xi)} \leq \int_\mathcal{X}\int_\mathcal{Y} \big|f(y)\big|^2 \mu_x(\mathrm{d}y)\xi(\mathrm{d}x) = \int_{\mathcal{X}\times\mathcal{Y}} \big|f(y)\big|^2 \mathrm{d}(\mathfrak{c}_\xi(\mu)\otimes\xi) = \big\|f\big\|^2_{L^2(\mu_\mathcal{Y})}.$$

The above also shows that the operator (14) $=: T[\mathfrak{c}_\xi(\mu)]$ is well-defined and bounded. To prove identification, note first that, since $L^2(\mu_\mathcal{Y})$ is separable (e.g. [14, Prop. 3.4.5]; recall that $\mathcal{Y}$ is Polish), there is a dense set $\mathcal{D} \coloneqq \{g_j \mid j \in J\} \subset L^2(\mu_\mathcal{Y})$ with J countable. Now take any $\mu, \tilde{\mu} \in \mathcal{M}_1^\xi(\mathcal{X}\times\mathcal{Y})$ with $T[\mathfrak{c}_\xi(\mu)] = T[\mathfrak{c}_\xi(\tilde{\mu})]$ (thus $L^2(\mu_\mathcal{Y}) = L^2(\tilde{\mu}_\mathcal{Y})$ in particular). Then for each $j \in J$, we have that $\mu_\cdot(g_j) = \tilde{\mu}_\cdot(g_j)$ in $L^2(\xi)$ and, thus, $\mu_x(g_j) = \tilde{\mu}_x(g_j)$ for each $x \in \mathcal{X}\setminus\mathcal{N}_j$ with $\mathcal{N}_j$ some ξ-nullset. In particular, for each $x \in \mathcal{X}\setminus\mathcal{N}$ with $\mathcal{N} \coloneqq \bigcup_{j\in J}\mathcal{N}_j$, we have $\mu_x|_\mathcal{D} = \tilde{\mu}_x|_\mathcal{D}$ and hence $\mu_x = \tilde{\mu}_x$ (since $\mathcal{D} \subset L^2(\mu_\mathcal{Y})$ is dense and each $\mathcal{M}_1(\mathcal{Y}) \ni \upsilon : L^2(\mu_\mathcal{Y}) \ni f \mapsto \int_\mathcal{Y} f \,\mathrm{d}\upsilon \in \mathbb{R}$ is bounded). Hence, as desired, $\mathfrak{c}_\xi(\mu) = \mathfrak{c}_\xi(\tilde{\mu})$ in $L^2(\xi)$, since $\xi(\mathcal{N}) = 0$ (as J is countable).

For the remaining identification of (14) with the function (16), take any $\mu, \tilde{\mu} \in \mathcal{M}_1^\xi(\mathcal{X}\times\mathcal{Y})$ with $\mu_\cdot(q) = \tilde{\mu}_\cdot(q)$ in $L^2(\xi; \mathcal{H}_\mathcal{Y})$ (i.e. such that $\mu_x(q) = \tilde{\mu}_x(q)$ for each $x \in \mathcal{X}\setminus\hat{\mathcal{N}}$ for some ξ-nullset $\hat{\mathcal{N}}$). Then for each $f \in \mathfrak{H}_q \coloneqq \{\langle \ell, q(\cdot)\rangle \mid \ell \in \mathcal{H}_\mathcal{Y}\}$, say $f = \langle \ell_f, q(\cdot)\rangle$, we have the $L^2(\xi)$-identities

$$\mathfrak{c}_\xi(\mu)(f) = \mu_\cdot\big(\langle \ell_f, q(\cdot)\rangle\big) = \langle \ell_f, \mu_\cdot(q)\rangle = \langle \ell_f, \tilde{\mu}_\cdot(q)\rangle = \mathfrak{c}_\xi(\tilde{\mu})(f), \quad \text{i.e.:} \quad \mathfrak{c}_\xi(\mu)|_{\mathfrak{H}_q} = \mathfrak{c}_\xi(\tilde{\mu})|_{\mathfrak{H}_q} \tag{124}$$

(where the second and the fourth identity holds since Bochner integrals commute with bounded functionals; e.g. [58, Sec. 1.3.1]). But since $\mathfrak{H}_q$ is dense in $L^2(\mu_\mathcal{Y})$ and the operator (14) is bounded, (124) implies that $\mathfrak{c}_\xi(\mu) = \mathfrak{c}_\xi(\tilde{\mu})$ (as operators (14) and thus, via the lemma's initial identification, also in the original kernel sense (9)) by uniqueness of extension.

Statement (17) is clear from the second equality in (124) and the assumption $L^2(\mu_\mathcal{Y}) = \overline{\mathfrak{H}_q}^{L^2}$. □

B.1.3 Proof of Lemma 2.9

Proof. Abbreviating the domain and range of (25) by H_1 and H_2 respectively, note that both $\iota_1 : L^2(\mu_\mathcal{Y}) \ni f \mapsto f\circ\hat{\pi}_2 \in H_1$ and $\iota_2 : L^2(\xi) \ni g \mapsto g\circ\hat{\pi}_1 \in H_2$ are well-defined Hilbert isometries and onto. Indeed: Well-definedness is clear since $\mu_\mathcal{Y} = (\hat{\pi}_2)_*\mu$ and $\xi = (\hat{\pi}_1)_*\mu$, and isometry holds since $\langle \iota_1(f), \iota_1(g)\rangle_{H_1} = \int_{\mathcal{X}\times\mathcal{Y}} (fg)\circ\hat{\pi}_2 \,\mathrm{d}\mu = \int_\mathcal{Y} fg \,\mathrm{d}[(\hat{\pi}_2)_*\mu] = \langle f, g\rangle_{L^2(\mu_\mathcal{Y})}$ (and likewise for ι_1); the surjectivity of ι_2 (and likewise that of ι_2) is seen as follows: if $\tilde{f} : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$

is Borel-measurable wrt. $\mathcal{X} \times \mathcal{B}(\mathcal{Y}) \equiv \{\mathcal{X} \times A \mid A \in \mathcal{B}(\mathcal{Y})\}$, then $\tilde{f}$ is constant wrt. its first variable (otherwise, i.e. if $\tilde{f}(x_1, y) \neq \tilde{f}(x_2, y)$ for a $y \in \mathcal{Y}$ and some $x_1, x_2 \in \mathcal{X}$ with $x_1 \neq x_2$, then: $A_{(x_1,y)} \coloneqq \tilde{f}^{-1}(\{\tilde{f}(x_1, y)\}) \cap (\mathcal{X} \cap \{y\}) \in \mathcal{X} \times \mathcal{B}(\mathcal{Y})$ (by measurability of $\tilde{f}$ and since all singletons in $\mathcal{Y}$ are (closed and thus) in $\mathcal{B}(\mathcal{Y})$) and $\hat{\pi}_1(A_{(x_1,y)}) \notin \{\emptyset, \mathcal{X}\}$ — a contradiction), that is: there is $f : \mathcal{Y} \to \mathbb{R}$ such that $\tilde{f}(x, y) = f(y)$ for each $(x, y) \in \mathcal{X} \times \mathcal{Y}$, which is equivalent to $\tilde{f} = f \circ \hat{\pi}_2$.

We now claim that the operator $\mathfrak{c}_\xi(\mu)$ from (14) relates to the operator (25) via

$$\mathfrak{c}_\xi(\mu) \;=\; \iota_2^{-1} \circ P_{\mathcal{Y}|\mathcal{X}} \circ \iota_1. \tag{125}$$

To verify this, we abbreviate $\tilde{T} \coloneqq \iota_2 \circ \mathfrak{c}_\xi(\mu) \circ \iota_1^{-1}$ and need to show that, for any $\tilde{f} \in H_1$,

$$\tilde{f} - \tilde{T}\tilde{f} \;\in\; H_2^\intercal, \quad \text{i.e.} \quad \big\langle \tilde{f} - \tilde{T}\tilde{f}, \tilde{g} \big\rangle_{L^2(\mu)} = 0 \quad \text{for each} \quad \tilde{g} \in H_2. \tag{126}$$

(Note here that the $P_{\mathcal{Y}|\mathcal{X}}$-defining projector P is well-defined since the orthogonal decomposition $L^2(\mu) = H_2 \oplus H_2^\intercal$ exists by the fact (consequent to every L^2-convergent subsequence admitting an μ-a.e. convergent subsequence) that H_2 is a closed subspace of $L^2(\mu)$.) Proving (126), note that

$$\begin{aligned}\int_{\mathcal{X}\times\mathcal{Y}} \mathbb{1}_{\tilde{A}} \Delta_{\tilde{f}} \,\mathrm{d}\mu \;&=\; \int_{A\times\mathcal{Y}} \tilde{f} \,\mathrm{d}\mu - \int_A \mu_x(f)\, \xi(\mathrm{d}x) \;=\; \int_{A\times\mathcal{Y}} \tilde{f} \,\mathrm{d}\mu - \int_{\mathcal{X}}\int_{\mathcal{Y}} f(y) \mathbb{1}_{\tilde{A}}(x, y)\, \mu_x(\mathrm{d}y)\, \xi(\mathrm{d}x) \\ &\overset{(7)}{=} \int_{A\times\mathcal{Y}} \tilde{f} \,\mathrm{d}\mu - \int_{\mathcal{X}\times\mathcal{Y}} \tilde{f}(x, y) \mathbb{1}_{\tilde{A}}(x, y)\, (\mu_{\mathcal{Y}|\mathcal{X}} \otimes \xi)(\mathrm{d}(x, y)) \overset{(8)}{=} 0 \qquad \big(f \coloneqq \iota_1^{-1}(\tilde{f})\big)\end{aligned}$$

for each $\tilde{A} \equiv (A \times \mathcal{Y}) \in \mathcal{B}(\mathcal{X}) \times \mathcal{Y}$ and with $\Delta_{\tilde{f}} \coloneqq \tilde{f} - \tilde{T}\tilde{f}$. This proves (126) and, hence, (125). □

B.2 Ad Section 3

B.2.1 Remarks on Stochastic Processes and Weighted Signature-Codomains

Remark B.3 (On Stochastic Processes)**.** 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 general facts on stochastic processes S as in (32).

(i) The rv X is $\big(\mathscr{F}, \mathcal{B}(\|\cdot\|_{1\text{-var}})\big)$-measurable by def. Now $\mathcal{B}(\|\cdot\|_{1\text{-var}}) = \mathcal{B}(\|\cdot\|_\infty)$ by Lemma 3.2, 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. [46, 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. [60, Proposition 2.23]), but for our purposes the weak notion of joint measurability will suffice, cf. also Lemma 3.5.

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

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

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. [23, 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 twins $\tilde{X} \coloneqq \mathbb{1}_{\tilde{\Omega}} \cdot X + \mathbb{1}_{\Omega\setminus\tilde{\Omega}} \cdot x_0 : \Omega \to \mathfrak{D}_X$ (any $x_0 \in \mathfrak{D}_X$ fixed, say $x_0 = 0$ for convenience) are indistinguishable.

(iv) We 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 $\mathcal{N}_\mathbb{P} \coloneqq \{N \subseteq \Omega \mid \exists\, A \in \mathscr{F} : N \subseteq A \text{ and } \mathbb{P}(A) = 0\}$ 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. [17, 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).

Remark B.4 (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 (40) 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 (34)], 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

$$\mathcal{H}_d^\gamma \coloneqq \Big\{ \boldsymbol{t} \in V \;\Big|\; \|\boldsymbol{t}\|_\gamma \coloneqq \sqrt{\textstyle\sum_{m\geq 0} \gamma_m \|\pi_m(\boldsymbol{t})\|_m^2} < \infty \Big\} \tag{127}$$

together with the inner product $\langle \boldsymbol{s}, \boldsymbol{t}\rangle_\gamma \coloneqq \sum_{m\geq 0} \gamma_m \langle \pi_m(\boldsymbol{s}), \pi_m(\boldsymbol{t})\rangle_m$. It is clear that $(\mathcal{H}_d^\gamma, \langle\cdot,\cdot\rangle)$ 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. [15, 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 $\|\cdot\|_{\tilde\gamma} \leq \|\cdot\|_\gamma$. ◆

B.2.2 Proof of Lemma 3.2

Proof. The first assertion holds by [18, Propositions 1.31 & 1.32]. In fact, [18, 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, [25, 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 $(\mathcal{C}^1, \|\cdot\|_{1\text{-var}})$ is separable and Banach is stated as [18, 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

$$B_r^1(x) \coloneqq \{y \in \mathcal{C}^1 \mid \|y - x\|_{1\text{-var}} \leq r\} \in \mathcal{B}_\infty \quad \text{for every} \quad x \in \mathcal{C}^1 \text{ and any } r \geq 0. \tag{128}$$

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 (128) holds as desired. □

B.2.3 Proof of Lemma 3.4

Proof. Note that since both $(\mathcal{X}, \|\cdot\|_{1\text{-var}})$ and $(\mathcal{Y}, \|\cdot\|_{1\text{-var}})$ are Polish by Lemma 3.2, 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

$$\mathcal{B}(\mathcal{Z}, \|\cdot\|_{1\text{-var}}) = \mathcal{B}(\mathcal{Z}, \|\cdot\|_\alpha), \tag{129}$$

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 $\|z\|_\beta := \max\{\|\pi_\mathcal{X}(z)\|_\infty, \|\pi_\mathcal{Y}(z)\|_\infty\}$ by Lemma 3.2 (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 (129) (upon recalling that $\mathcal{B}(\mathcal{Z}, \|\cdot\|_\alpha) = \mathcal{B}(\mathcal{X}, \|\cdot\|_{1\text{-var}}) \otimes \mathcal{B}(\mathcal{Y}, \|\cdot\|_{1\text{-var}})$) and the fact that a product-space-valued function (here: (X, Y)) is product-measurable iff all of its factor components (here: X and Y) are measurable. □

B.2.4 Proof of Lemma 3.12

Proof. Each of the spaces $(V_m, \langle\cdot,\cdot\rangle_m)$ from (40) is Hilbert and separable (with ONB $[d]^*_m := \{w \in [d]^* \mid |w| = m\}$). Thus the space $\mathcal{H}$ is Hilbert and separable — with ONB $[d]^*$ — as the Hilbert direct sum of the family $\{(V_m, \langle\cdot,\cdot\rangle_m) \mid m \in \mathbb{N}_0\}$, see e.g. [15, Proposition I.6.2]. The inclusion $\mathfrak{sig}(\mathcal{C}^1_d) \subset \mathcal{H}$ follows from the factorial decay of the signature coefficients, cf. [11, Corollary 5.5].

The last assertion follows from the usual p-variation continuity of $\mathfrak{sig}$, see e.g. [11, Corollary 5.5], and the fact that the locally convex topology from [11, 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. [11, Corollary 2.4], the topological space (V, τ_{lc}) is sequential, whence $\tau_\mathcal{H} \subseteq \tau_{\mathrm{lc}}$ 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.2.5 Proof for Remark 3.16

Claim: *On $\mathcal{H}^\downarrow_d$, the locally convex topology $\tau_\downarrow$ is* finer *than the* (127)*-induced subspace topology τ_γ.*

Proof. Since $\tau_\downarrow$ is metrizable (Lemma 3.15), the space $(\mathcal{H}^\downarrow_d, \tau_\downarrow)$ is sequential, whence $\tau_\gamma \subseteq \tau_\downarrow$ (iff $\mathrm{id}_{\mathcal{H}^\downarrow_d} : (\mathcal{H}^\downarrow_d, \tau_\downarrow) \to (\mathcal{H}^\downarrow_d, \tau_\gamma)$, $v \mapsto v$, is continuous) iff every $\tau_\downarrow$-convergent sequence in $\mathcal{H}^\downarrow_d$ is τ_γ-convergent. This clearly holds, however, since for every null-sequence $(\boldsymbol{v}_k)$ in $(\mathcal{H}^\downarrow_d, \tau_\downarrow)$ there is $k_0 \in \mathbb{N}$ with $\sup_{k\geq k_0} |||\boldsymbol{v}_k|||_\lambda < 1$ and hence, for each $k \geq k_0$,

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

As one consequence of the inclusion $\tau_\gamma \subseteq \tau_\downarrow$, statement (45) also holds for $\mathcal{H}^\downarrow_d$ replaced by $\mathcal{H}^\gamma_d$. □

B.2.6 Proof of Lemma 3.17

Proof. Since the augmentation map $\bar{\iota} = (\theta, \mathrm{id}_\mathcal{X}) : \mathcal{C}^1_d \to \mathcal{C}^1_{\underline{d}}$ from (38), with $\theta(x) := (t)_{t\in[0,1]}$, is $(\|\cdot\|_{p\text{-var}}, \|\cdot\|_{p\text{-var}})$-continuous, the continuity (47) follows immediately from (45) and assertion (46).

Let us now prove (46). Recall for this that $\Lambda(\boldsymbol{t}) = \sum_{m\geq 0} \lambda^m_{\boldsymbol{t}} \pi_m(\boldsymbol{t})$ by definition, and that, since $(\mathcal{H}^\downarrow_d, \tau_\downarrow)$ is first-countable by Lemma 3.15, the map Λ is continuous if it is sequentially continuous.

Let hence $\boldsymbol{t}, (\boldsymbol{t}_k) \subset \mathcal{H}_d^{\downarrow}$ 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

$$\|\Lambda(\boldsymbol{t}_k) - \Lambda(\boldsymbol{t})\|^2 \;\le\; 2 \sum_{m\ge 0} \big(\|\lambda^m_{\boldsymbol{t}_k} \pi_m(\boldsymbol{t}_k) - \lambda^m_{\boldsymbol{t}_k} \pi_m(\boldsymbol{t})\|_m^2 + \|\lambda^m_{\boldsymbol{t}_k} \pi_m(\boldsymbol{t}) - \lambda^m_{\boldsymbol{t}} \pi_m(\boldsymbol{t})\|_m^2 \big). \tag{130}$$

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, $\lim_{k\to\infty} |||\boldsymbol{t}_k - \boldsymbol{t}|||_c = 0$ and $|||\boldsymbol{t}|||_c < \infty$. The summands $\alpha_{m,k} := \|\lambda^m_{\boldsymbol{t}_k} \pi_m(\boldsymbol{t}_k - \boldsymbol{t})\|_m^2$ and $\beta_{m,k} := \|(\lambda^m_{\boldsymbol{t}_k} - \lambda^m_{\boldsymbol{t}})\pi_m(\boldsymbol{t})\|_m^2$ on the right-hand side of (130) compare to

$$\alpha_{m,k} \;\le\; a_{m,k} := c^m \|\pi_m(\boldsymbol{q}_k - \boldsymbol{q})\|_m \quad \text{and} \quad \beta_{m,k} \;\le\; b_m := 4c^m \|\pi_m(\boldsymbol{q})\|_m, \tag{131}$$

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 (130) we find

$$\lim_{k\to\infty} \|\Lambda(\boldsymbol{t}_k) - \Lambda(\boldsymbol{t})\|^2 \;\le\; 2 \lim_{k\to\infty} |||\boldsymbol{t}_k - \boldsymbol{t}|||_c + \sum_{m\ge 0} \lim_{k\to\infty} \beta_{m,k} \;=\; 0\,, \tag{132}$$

where interchanging limit and summation for the second summand in (132) is permissible by dominated convergence, which in turn is applicable thanks to the $(\beta_{m,k})$-domination in (131) and the fact that $\sum_{m\ge 0} |b_m| = 4|||\boldsymbol{q}|||_c < \infty$. This proves (46), as desired. □

B.2.7 Definition of the Strict Topology and Detailed Proof of Proposition 3.19

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

$$B_0(\mathcal{Z}) := \{\psi : \mathcal{Z} \to \mathbb{R} \text{ bounded} \mid \forall\, \varepsilon > 0 \,:\, \exists\, \mathcal{K} \subset \mathcal{Z} \text{ compact} \;:\; \sup\nolimits_{x\in\mathcal{Z}\setminus\mathcal{K}} |\psi(x)| < \varepsilon\}$$

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

Definition B.5 (Strict Topology [19])**.** The *strict topology* on $C_b(\mathcal{Z})$, denoted by $\tau^{\mathcal{Z}}_{\mathrm{str}}$, is the topology induced by the family of seminorms

$$p_\psi(f) \;:=\; \sup_{x\in\mathcal{Z}} \big|f(x)\psi(x)\big|, \quad \psi \in B_0(\mathcal{Z}).$$

Note that the strict topology is weaker than the uniform topology on $C_b(\mathcal{Z})$ but stronger than the topology of compact (and thus also pointwise) convergence on $C_b(\mathcal{Z})$, see [19, Theorem 2.4 (i)].

Proof of Proposition 3.19. Let us first note that, indeed,

$$\sum\nolimits_{w\in[\underline{d}]^*} \underline{\xi}^{\Lambda}_w \cdot w = \Lambda \circ \underline{\mathfrak{sig}}. \tag{133}$$

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 $(x, m) \in \mathcal{Z} \times \mathbb{N}_0$.

To see $\mathcal{A}_\Lambda \subset C(\mathcal{Z})$, note [from (50)] that since each function $\varphi \in \mathcal{A}_\Lambda$ can be represented as

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

the desired $\|\cdot\|_{1\text{-var}}$-continuity of φ follows from (133) and statement (47) of Lemma 3.17.

Since for $\mathbf{1} = 1 \cdot \emptyset \in \mathbb{R}[\underline{d}]$ 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 $\mathcal{Z}$, clearly $\mathcal{A}_\Lambda$ is non-vanishing. For $\mathcal{A}_\Lambda$ being point-separating, note that for any $x, y \in \mathcal{Z}$ with $x \ne y$ we have $\underline{\mathfrak{sig}}(x) \ne \underline{\mathfrak{sig}}(y)$ by Lemma 39, and hence also $\underline{\mathfrak{sig}}_\Lambda(x) \ne \underline{\mathfrak{sig}}_\Lambda(y)$ by (133) and the injectivity of Λ. This implies that there is $w_0 \in [\underline{d}]^*$ such that $\langle w_0, \underline{\mathfrak{sig}}_\Lambda(x)\rangle \ne \langle w_0, \underline{\mathfrak{sig}}_\Lambda(y)\rangle$, whence for $\varphi := \underline{\xi}^{\Lambda}_{w_0} \in \mathcal{A}_\Lambda$ we find $\varphi(x) \ne \varphi(y)$.

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,

$$\begin{aligned}
\varphi \cdot \psi &= \sum_{w,\tilde{w} \in [\underline{d}]^*} \langle \ell_\varphi, w \rangle \langle \ell_\psi, \tilde{w} \rangle \underline{\lambda}_\cdot^{|w|} \underline{\lambda}_\cdot^{|\tilde{w}|} \langle w, \underline{\mathfrak{sig}} \rangle \langle \tilde{w}, \underline{\mathfrak{sig}} \rangle && (134)\\
&= \sum_{w,\tilde{w} \in [\underline{d}]^*} \underline{\lambda}_\cdot^{|w|+|\tilde{w}|} \langle \ell_\varphi, w \rangle \langle \ell_\psi, \tilde{w} \rangle \langle w \sqcup\!\sqcup \tilde{w}, \underline{\mathfrak{sig}} \rangle && (135)\\
&= \sum_{m \geq 0} \underline{\lambda}_\cdot^m \sum_{w,\tilde{w} \,:\, |w \sqcup\!\sqcup \tilde{w}| = m} \langle \ell_\varphi, w \rangle \langle \ell_\psi, \tilde{w} \rangle \langle w \sqcup\!\sqcup \tilde{w}, \underline{\mathfrak{sig}} \rangle && (136)\\
&= \sum_{m \geq 0} \underline{\lambda}_\cdot^m \langle \pi_m(\ell_{\varphi\psi}), \underline{\mathfrak{sig}} \rangle && (137)\\
&= \langle \ell_{\varphi\psi}, \Lambda \circ \underline{\mathfrak{sig}} \rangle \,, \quad \text{for} \quad \ell_{\varphi\psi} := \sum\nolimits_{w,\tilde{w} \in [\underline{d}]^*} \langle \ell_\varphi, w \rangle \langle \ell_\psi, \tilde{w} \rangle \cdot w \sqcup\!\sqcup \tilde{w}. && (138)
\end{aligned}$$

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

While (134) holds simply by linearity of (41), equation (135) involved the *character identity*

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

(see e.g. [36, proof of Thm. 2.15]), where $\sqcup\!\sqcup : \mathbb{R}[\underline{d}]^{\times 2} \to \mathbb{R}[\underline{d}]$ is the so-called *shuffle product*, defined e.g. in [36, eq. (2.5) (p. 35)]. Denoting by $|\ell| := \max\{|w| \mid w \in [\underline{d}]^* : w \in \ell\}$ the maximal length of any word contained (as a summand) in a given polynomial $\ell \in \mathbb{R}[\underline{d}]$, it holds that $|\ell_1 \sqcup\!\sqcup \ell_2| = |\ell_1| + |\ell_2|$. This justifies equation (136), where we also used that $[\underline{d}]^* \times [\underline{d}]^* = \bigsqcup_{m \geq 0} \{(w, \tilde{w}) \in [\underline{d}]^* \times [\underline{d}]^* \mid |w \sqcup\!\sqcup \tilde{w}| = m\}$ defines a (disjoint) partition. For equation (137) 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 (138) 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 (41) of $\langle \cdot, \cdot \rangle$).

We thus saw that $\mathcal{A}_\Lambda$ is a subalgebra of $C(\mathcal{Z})$. Now if Λ is in fact an fN of the form (48), then

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

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

The asserted denseness of $\mathcal{A}_\Lambda$ in $(C_b(\mathcal{Z}), \tau_{\mathrm{str}}^{\mathcal{Z}})$ is then guaranteed by [19, Theorem 3.1], which generalises the theorem of Stone-Weierstrass to the $\tau_{\mathrm{str}}^{\mathcal{Z}}$-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 $C(\mathcal{Z})$, while the bounds (44) yield that, for each $\varphi \in \mathcal{A}$,

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

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

The following corollary is an immediate consequence of Proposition 3.19.

Corollary B.6. *Given $d \in \mathbb{N}$, let $\mathcal{Z}$ be a subset of $\mathcal{C}_d^1$. Then for each $f \in C_b(\mathcal{Z})$ we have*

$$f \;=\; \lim_{k \to \infty} \big\langle \tilde{\ell}_k, \Lambda \circ \underline{\mathfrak{sig}} \big\rangle \;\; \textit{in} \;\; \tau_{\mathrm{str}}^{\mathcal{X}}, \quad \textit{for some} \quad (\tilde{\ell}_k) \subset \mathbb{R}[\underline{d}].$$

B.2.8 Proof of Lemma 4.3

Proof. Let us abbreviate $\varphi \coloneqq \underline{\mathfrak{sig}}_{\underline{\Xi}}$ and $\phi_w \coloneqq \underline{\xi}^{\Lambda}_w(X)$ for each $w \in [\underline{d}]^*$. We first prove that $\sigma(X) = \sigma(\varphi(X))$: The inclusion $\sigma(\varphi(X)) \subseteq \sigma(X)$ is immediate since $\varphi : \mathcal{X} \to \mathcal{H}_{\underline{d}}$ is continuous (and hence Borel-measurable), see Lemma 3.17. For the converse, note that since φ is also an injection, we have that $A_\varphi \coloneqq \varphi(A) \in \mathcal{B}(\mathcal{H}_{\underline{d}}) \cap \varphi(\mathcal{X})$ for any fixed $\|\cdot\|_{1\text{-var}}$-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 $A^c_\varphi \equiv \varphi(\mathcal{X}) \setminus A_\varphi = \varphi(A^c)$, where the last identity holds since φ is injective; this implies that $A_\varphi \in \mathcal{B}(\mathcal{H}_{\underline{d}} \cap \varphi(\mathcal{X}))$ by a theorem of Souslin [28, 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 (56).

The inclusion $\sigma(\phi_w \mid w \in [\underline{d}]^*) \subseteq \sigma(\varphi(X))$ is immediate since $\phi_w = \langle w, \varphi(X)\rangle$ for each $w \in [\underline{d}]^*$ (cf. (41)) and each $\langle w, \cdot\rangle : \mathcal{H}_{\underline{d}} \to \mathbb{R}$ is continuous. The converse inclusion holds as, pointwise on Ω,

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

(cf. (50) and (47)), i.e. $\varphi(X)$ is the pointwise limit of $\sigma(\phi_w|\, w)$-measurable functions and thus $\sigma(\phi_w|\, w)$-measurable itself. □

B.2.9 Proof of Theorem 4.8

Proof. We begin with the well-known observation that the space $L^2_X(\mathcal{H}_\mathcal{Y})$ from (57) 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

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

(The closedness of $L^2_X(\mathcal{H}_\mathcal{Y})$ follows from the well-known fact (which persists for Hilbert-valued random variables [58, Proposition 2.11]) that L^2-convergence implies almost sure convergence on a subsequence.) Denoting $E \coloneqq L^2(\mathbb{P};\mathcal{H}_\mathcal{Y})$ and $G \coloneqq L^2_X(\mathcal{H}_\mathcal{Y})$ for brevity, we then adopt (from the scalar-valued setting, e.g. [58, Section 11.1]) the classical perspective that the (vector-valued) conditional expectation $\mathbb{Y}^\Lambda_X \coloneqq \mathbb{E}\big[\mathbb{Y}^\Lambda \,\big|\, X\big] \in G$ is the orthogonal projection of $\mathbb{Y}^\Lambda$ onto G along $G^\perp$, in symbols:

$$\mathbb{Y}^\Lambda_X \;=\; \mathrm{P}_G \mathbb{Y}^\Lambda \quad \text{for the orthogonal projector } \;\mathrm{P}_G : E \to E \;\text{ on }\; G\,(=\mathrm{im}(\mathrm{P}_G)). \tag{139}$$

To see that this perspective (139) is true in the present vector-valued setting (54) & (55), denote $\mathbb{Y}_G \coloneqq \mathrm{P}_G\mathbb{Y}^\Lambda$ and notice that then $\Delta \coloneqq \mathbb{Y}^\Lambda - \mathbb{Y}_G \in G^\perp$, 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. [58, Sec. 1.3.1]; note that each element of E is Bochner-integrable by definition (54) and [58, 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. [58, Theorem 11.10], which implies that $\mathbb{Y}_G = \mathbb{Y}^\Lambda_X$ as claimed in (139).

The above characterisation (139) of $\mathbb{Y}^\Lambda_X$ as the orthogonal projection of $\mathbb{Y}^\Lambda$ onto G implies that

$$\big\|\mathbb{Y}^\Lambda - \mathbb{Y}^\Lambda_X\big\|_{L^2(\mathcal{H}_\mathcal{Y})} \le\; \big\|\mathbb{Y}^\Lambda - \mathbb{Z}\big\|_{L^2(\mathcal{H}_\mathcal{Y})} \quad \text{for all } \;\mathbb{Z} \in G. \tag{140}$$

Moreover, the Hilbert projection theorem guarantees that the arg min in (140) is unique, so that in fact

$$\mathbb{Y}_X^\Lambda \;=\; \operatorname*{arg\,min}_{\mathbb{Z}\in G} \mathbb{E}\big\|\mathbb{Y}^\Lambda - \mathbb{Z}\big\|^2. \tag{141}$$

Now in order to make the variational identity (141) more operational, recall from Prop. 4.5 that

$$\text{the set}\;\; \mathfrak{G} \coloneqq \Psi(X) = \big\{\psi_\alpha(X)\,\big|\,\alpha\in\mathfrak{L}_X^2\big\}\;\; \text{from}\;\; (59)\;\; \text{is}\;\; \|\cdot\|_{L^2(\mathcal{H}_\mathcal{Y})}\text{-dense in}\;\; G. \tag{142}$$

Let us observe how (142) and (141) imply (65): Abbreviating $\Phi(\mathbb{Z}) \coloneqq \|\mathbb{Y}^\Lambda - \mathbb{Z}\|^2_{L^2(\mathcal{H}_\mathcal{Y})}$, 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

$$\Phi(\mathbb{Y}_X^\Lambda) \overset{(141)}{=} \inf_{\mathbb{Z}\in G}\Phi(\mathbb{Z}) \overset{(142)}{=} \inf_{\mathbb{Z}\in\mathfrak{G}}\Phi(\mathbb{Z}) \;=\; \inf_{\alpha\in\mathfrak{L}_X^2}\Phi(\psi_\alpha(X)) \;=:\; \gamma.$$

Hence for any minimizing sequence of (66), i.e. any sequence (α_k) in $\mathfrak{L}_X^2$ with $\lim_{k\to\infty}\Phi(\psi_{\alpha_k}(X)) = \gamma$, the functions $(\mathbb{Z}_k) \coloneqq (\psi_{\alpha_k}(X)) \subset G$ 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

$$\|\mathbb{Z}_n - \mathbb{Z}_m\|^2_{L^2(\mathcal{H}_\mathcal{Y})} \;\le\; 2\Phi(\mathbb{Z}_n) + 2\Phi(\mathbb{Z}_m) - 4\Phi(\mathbb{Y}_X^\Lambda) \;\;\longrightarrow\;\; 0 \quad (\text{for}\;\; n,m\to\infty), \tag{143}$$

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 $\Phi(\mathbb{Z}_\star) = \lim_{k\to\infty}\Phi(\mathbb{Z}_k) = \gamma$ and thus

$$\mathbb{Z}_\star \in \operatorname*{arg\,min}_{\mathbb{Z}\in G}\Phi(\mathbb{Z}), \quad \text{which, by (141), implies} \quad \mathbb{Z}_\star = \mathbb{Y}_X^\Lambda$$

(as noted in the lead-up to (140), the above arg min contains exactly one element only). This proves (65). Regarding the final claim on almost sure convergence, note $\|\mathbb{Z}_k - \mathbb{Z}_\star\|^2_{L^1(\mathcal{H}_\mathcal{Y})} \le \|\mathbb{Z}_k - \mathbb{Z}_\star\|^2_{L^2(\mathcal{H}_\mathcal{Y})} \le 2(\Phi(\mathbb{Z}_k) - \gamma) =: 2\beta_k$ by (143) [and since $\|\cdot\|_{L^1} \le \|\cdot\|_{L^2}$], 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. □

B.2.10 Proof of Corollary 4.11

Proof. Let $f \in C(\mathfrak{D}_Y)$. As $\mathfrak{D}_Y \coloneqq \operatorname{supp}(\mathbb{P}_Y)$ is assumed compact, there is $(\ell_f^{(\nu)})_\nu \subset \mathbb{R}[\tilde{d}]$ such that

$$\lim_{\nu\to\infty}\big\|f - \langle\ell_f^{(\nu)}, \underline{\mathfrak{sig}}_\Lambda\rangle\big\|_{\infty;\mathfrak{D}_Y} \;=\; 0 \tag{144}$$

by Corollary B.6, so minimizing sequences for the optimisation problem (76) exist.

Since $\mathbb{P}(Y\in\mathfrak{D}_Y) = 1$, or even $Y(\Omega)\subseteq\mathfrak{D}_Y$ without loss of generality by Remark B.3 (iii),

$$C \coloneqq \sup_{\mu\in\mathbb{N}}\sup_{\omega\in\Omega}\big|\varsigma_\mu(Y(\omega))\big| \;<\; \infty \quad \text{for} \quad \varsigma_\nu(y) \coloneqq f(y) - \langle\ell_f^{(\nu)}, \underline{\mathfrak{sig}}_\Lambda(y)\rangle.$$

Indeed, given $\epsilon > 0$ there is $\nu_0\in\mathbb{N}$ with $\sup_{\nu\ge\nu_0,\omega}|\varsigma_\nu(Y(\omega))| \le \sup_{\nu\ge\nu_0}\|\varsigma_\nu\|_{\infty;\mathfrak{D}_Y} \le \epsilon$ by (144), and so $C \le \max_{\nu\le\nu_0}\|\varsigma_\nu\|_{\infty;\mathfrak{D}_Y} + \epsilon \le 1 + \|\Lambda\|_{\infty;\mathcal{H}}\cdot\max_{\nu\le\nu_0}\|\ell_f^{(\nu)}\| + \epsilon < \infty$. Thus and with (144),

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

via the conditional dominated convergence theorem, which proves (α) in (77). The convergence (β) follows as in (73), while the corollary's last assertion is immediate from (77) and (65). □

B.3 Ad Section 5

B.3.1 Proof of Proposition 5.1

Proof. That $\mathfrak{H}_\Xi$ is a vRKHS with reproducing kernel K is shown in [7, Example 3.2 (iii)]. This reference also implies that $\mathfrak{H}_\Xi$ is separable: Denoting by $\varphi \coloneqq \underline{\mathfrak{sig}}_\Xi : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ the feature map of $\mathcal{H}_\kappa$, we see (e.g. from [40, Lemma 2.1]) that the RKHS $\mathcal{H}_\kappa$ is separable due to the separability of $\varphi(\mathcal{X})$, where the latter holds by the separability of $\mathcal{X}$ (Lemma 3.2) and the fact that φ is continuous (Lemma 3.17). The asserted separability of $\mathfrak{H}_\Xi$ now follows from the unitary equivalence $\mathfrak{H}_\Xi \cong \bigoplus_{w\in[\tilde{d}]^*} \mathcal{H}_\kappa$ (given in [7, Example 3.2 (iii)]) and the fact that the countable direct sum of separable Hilbert spaces is separable.

For a proof of (81), note first that $\mathfrak{H}_\Xi$ can be canonically injected into $L^2(\mathcal{X}, \mathbb{P}_X; \mathcal{H}_\mathcal{Y})$ via [42, Lemma 2.1 (i)] by the fact that $\sup_{x\in\mathcal{X}} \|K(x,x)\|_{\mathrm{op}} = |\kappa(x,x)| \leq \sup_{\boldsymbol{t}\in\mathcal{H}_\mathcal{Y}} \|\Lambda(\boldsymbol{t})\|^2_{\mathcal{H}_\mathcal{Y}} =: R^2_\Lambda < \infty$. (The Borel-measurability of each $f \in \mathfrak{H}_\Xi$ holds since $(w \mid w \in [\tilde{d}]^*)$ is a countable ONB of $\mathcal{H}_\mathcal{Y}$ and the feature map $\underline{\mathfrak{sig}}_\Xi$ of $\mathcal{H}_\kappa$ is continuous.) Next, take any $g \in L^2(\mathcal{X}, \mathbb{P}_X; \mathcal{H}_\mathcal{Y})$ and $\varepsilon > 0$. Then

$$\sum_{m=0}^{\infty} \beta_m(g) \;=\; \|g\|^2_{L^2(\mathbb{P}_X;\mathcal{H}_\mathcal{Y})} \;<\; \infty \qquad \text{for} \quad \beta_m(g) \coloneqq \sum\nolimits_{|w|=m} \int_\mathcal{X} |\langle g(x), w\rangle|^2 \, \mathbb{P}_X(\mathrm{d}x)$$

and hence $\sum_{m=m_0+1}^{\infty} \beta_m(g) \leq \varepsilon^2/2$ for some $m_0 \in \mathbb{N}$. As done in (62), we can for all $w \in [\tilde{d}]^*$ find

$$\alpha_w \;\in\; \mathbb{R}[d] \quad \text{such that} \quad \big\|\langle g, w\rangle - \langle \alpha_w, \underline{\mathfrak{sig}}_\Xi\rangle\big\|^2_{L^2(\mathbb{P}_X)} = \big\|\langle g(X), w\rangle - \varphi_{\alpha_w}\big\|^2_{L^2(\mathbb{P})} \;\leq\; \varepsilon^2 (2\tilde{d})^{-|w|}/4$$

by the density (59) (resp. (61)). Hence for $f_\varepsilon \coloneqq \sum_{m=0}^{m_0} \sum_{|w|=m} \langle \alpha_w, \underline{\mathfrak{sig}}_\Xi\rangle w \in \mathfrak{H}_\Xi$ we have that

$$\big\|g - f_\varepsilon\big\|^2_{L^2(\mathbb{P}_X;\mathcal{H}_\mathcal{Y})} \;=\; \sum_{|w|\leq m_0} \big\|\langle g - f_\varepsilon, w\rangle\big\|^2_{L^2(\mathbb{P}_X)} + \sum_{m=m_0+1}^{\infty} \beta_m(g) \;\leq\; \varepsilon^2 \big(\textstyle\sum_{m=0}^{\infty} 2^{-m} + 2\big)/4 = \varepsilon^2$$

(cf. (63)). Since $g \in L^2(\mathcal{X}, \mathbb{P}_X; \mathcal{H}_\mathcal{Y})$ and $\varepsilon > 0$ have been arbitrary, statement (81) follows.

The proposition's subsequent assertion is an immediate corollary to the above argument.

For the remaining statements on (82), note first that clearly $\Psi^\star_X \subseteq L^2(\mathcal{X}, \mathbb{P}_X; \mathcal{H}_\mathcal{Y})$ (the arguments are the same as for the inclusion $\Psi_X \subseteq L^2(\mathbb{P}_X; \mathcal{H}_\mathcal{Y})$) and $\Psi_X \subseteq \Psi^\star_X$. Further, for any $f \in \mathfrak{H}_\Xi$ we have

$$\textstyle\sum_{w\in[\tilde{d}]^*} \mathbb{E}\big[\langle f(X), w\rangle^2\big] = \mathbb{E}\big[\|f(X)\|^2_{\mathcal{H}_\mathcal{Y}}\big] = \|f\|^2_{L^2(\mathbb{P}_X;\mathcal{H}_\mathcal{Y})} \;\leq\; R^2_\Lambda \|f\|^2_{\mathfrak{H}_\Xi} \;<\; \infty \tag{145}$$

by [42, Lemma 2.1 (i)], which shows that the inclusion $\mathfrak{H}_\Xi \subseteq L^2(\mathcal{P}_X; \mathcal{H}_\mathcal{Y})$ is continuous. Upon recalling that $\mathcal{H}_\kappa = \{\langle u, \underline{\mathfrak{sig}}_\Xi\rangle \mid u \in \mathcal{H}_\mathcal{X}\}$ (which is the feature-map representation of $\mathcal{H}_\kappa$), the estimate (145) also shows $\mathfrak{H}_\Xi \subseteq \Psi^\star_X$. That all inclusions are dense has been shown above. □

B.3.2 Proof of Lemma 5.2

Proof. Since the functions $\tilde{\phi}_{f,\lambda}$ and $\boldsymbol{\phi}^{(m)}_{\mu|\lambda}$ are strictly convex and coercive, they both admit a unique minimizer over their respective domains (proving the assertions on $\Upsilon^*_{\mu,\lambda,m}$), see e.g. the proof of [42, Lemma 2.4] for the full argument. In particular, $\tilde{\phi}_{f,\lambda}(h_\star) = \min_{h\in\mathcal{H}_{\tilde{\kappa}}} \tilde{\phi}_{f,\lambda}(h) =: \gamma_\star$ for some (unique) $h_\star \in \mathcal{H}_{\tilde{\kappa}}$. But since $\mathcal{H}_{\tilde{\kappa}}$ is an RKHS with feature map $\underline{\mathfrak{sig}}_\Lambda$, we have that (a) $\min_{\ell\in\mathcal{H}_\mathcal{Y}} \boldsymbol{\phi}^f_{\mu_\mathcal{Y}|\lambda}(\ell) = \gamma_\star$ (cf. the definition (83)), and (b) there is $\ell_\star \in \mathcal{H}_\mathcal{Y}$ such that $h_\star = \langle \ell_\star, \underline{\mathfrak{sig}}_\Lambda\rangle$. Hence and since $\phi_{f,\lambda}(h_\star) = \boldsymbol{\phi}^f_{\mu_\mathcal{Y}|\lambda}(\ell_\star)$, a minimizer $\ell^f_{\mu_\mathcal{Y},\lambda}$ as in (86) exists.

Suppose now that $\mu_{\mathcal{Y}}$ is finitely supported, i.e. a convex combination of the form $\mu_{\mathcal{Y}} = \sum_{i\in I} p_i \delta_{y_i}$ for some $\{y_i \mid i \in I\} \subset \mathcal{Y}$ with I finite (this representation is valid since the path space $\mathcal{Y}$ is Polish and hence Radon), and note that the associated function (83) then reads

$$\mathcal{H}_{\mathcal{Y}} \ni \ell \;\longmapsto\; \boldsymbol{\phi}^f_{\mu_{\mathcal{Y}}|\lambda}(\ell) = \sum_{i\in I} p_i \big| f(y_i) - \langle \ell, q_i\rangle\big)\big|^2 + \lambda \|q^*_\ell\|^2_{\tilde{\kappa}} \qquad \text{with} \quad q_i \coloneqq \underline{\mathfrak{sig}}_\Lambda(y_i) \tag{146}$$

and $q \coloneqq \underline{\mathfrak{sig}}_\Lambda : \mathcal{Y} \to \mathcal{H}_{\mathcal{Y}}$ and $q^*_\ell \coloneqq \langle \ell, q\rangle \in \mathcal{H}_{\tilde{\kappa}}$. Let further $q^*_i \coloneqq \langle q_i, q\rangle \in \mathcal{H}_{\tilde{\kappa}}$ for each $i \in I$. The derivation of (88) follows the idea behind the classical representer theorem [51], for which we consider the (closed) subspaces $V_{\mu_{\mathcal{Y}}} = \operatorname{span}\{q_i \mid i \in I\} \subset \mathcal{H}_{\mathcal{Y}}$ and $V^*_{\mu_{\mathcal{Y}}} \coloneqq \operatorname{span}\{q^*_i \mid i \in I\} \subset \mathcal{H}_{\tilde{\kappa}}$. Then $\mathcal{H}_{\mathcal{Y}} = V_{\mu_{\mathcal{Y}}} \oplus V^\perp_{\mu_{\mathcal{Y}}}$ and $\mathcal{H}_{\tilde{\kappa}} = V^*_{\mu_{\mathcal{Y}}} \oplus \big(V^*_{\mu_{\mathcal{Y}}}\big)^\perp$, so that for any fixed minimizer $\hat{\ell} \in \mathcal{H}_{\mathcal{Y}}$ of (146) we have $\hat{\ell} = \hat{\ell}_1 + \hat{\ell}_2$ and $q^*_{\hat{\ell}} = \hat{q}^*_1 + \hat{q}^*_2$ for some pairs $(\hat{\ell}_1, \hat{\ell}_2) \in V_{\mu_{\mathcal{Y}}} \times V^\perp_{\mu_{\mathcal{Y}}}$ and $(\hat{q}^*_1, \hat{q}^*_2) \in V^*_{\mu_{\mathcal{Y}}} \times \big(V^*_{\mu_{\mathcal{Y}}}\big)^\perp$ which are both unique with these properties; said uniqueness implies that $\hat{q}^*_\nu = \langle \hat{\ell}_\nu, q\rangle$ for $\nu = 1, 2$ (note for this that $\langle \hat{\ell}_2, q\rangle \in \big(V^*_{\mu_{\mathcal{Y}}}\big)^\perp$ since $\big\langle q^*_i, \langle \hat{\ell}_2, q\rangle\big\rangle_{\tilde{\kappa}} = \big\langle \tilde{\kappa}(y_i, \cdot), \langle \hat{\ell}_2, q\rangle\big\rangle_{\tilde{\kappa}} = \langle \hat{\ell}_2, q(y_i)\rangle = 0$ for each $i \in I$, by the reproducing property). Clearly $\|\hat{\ell}\|_{\mathcal{H}_{\mathcal{Y}}} \geq \|\hat{\ell}_1\|_{\mathcal{H}_{\mathcal{Y}}}$, and by orthogonality,

$$\langle \hat{\ell}, q_i\rangle = \langle \hat{\ell}_1, q_i\rangle \quad \text{for each } i \in I \quad \text{and} \quad \big\|q^*_{\hat{\ell}}\big\|^2_{\tilde{\kappa}} = \big\|q^*_{\hat{\ell}_1}\big\|^2_{\tilde{\kappa}} + \big\|\hat{q}^*_2\big\|^2_{\tilde{\kappa}} \geq \big\|\langle \hat{\ell}_1, q\rangle\big\|^2_{\tilde{\kappa}}. \tag{147}$$

Combining (146) and (147) implies $\boldsymbol{\phi}^f_{\mu_{\mathcal{Y}}|\lambda}(\hat{\ell}) \geq \boldsymbol{\phi}^f_{\mu_{\mathcal{Y}}|\lambda}(\hat{\ell}_1)$ and hence, as $\hat{\ell}$ is a minimizer of $\boldsymbol{\phi}^f_{\mu_{\mathcal{Y}}|\lambda}$, that

$$\boldsymbol{\phi}^f_{\mu_{\mathcal{Y}}|\lambda}(\hat{\ell}_1) \;=\; \min_{\ell\,\in\,\mathcal{H}_{\mathcal{Y}}} \boldsymbol{\phi}^f_{\mu_{\mathcal{Y}}|\lambda}(\ell) \;=\; \min_{\ell\,\in\,V_{\mu_{\mathcal{Y}}}} \boldsymbol{\phi}^f_{\mu_{\mathcal{Y}}|\lambda}(\ell).$$

Since the above minimizer $\hat{\ell}$ of (146) was arbitrary, the claims surrounding (88) follow for $\ell^{f,\star}_{\mu_{\mathcal{Y}},\lambda} \coloneqq \hat{\ell}_1$ if we can show that the function (146) has only one minimizer over $V_{\mu_{\mathcal{Y}}}$, i.e. that

$$\left|\operatorname*{arg\,min}_{\ell\,\in\,V_{\mu_{\mathcal{Y}}}} \boldsymbol{\phi}^f_{\mu_{\mathcal{Y}}|\lambda}(\ell)\right| = 1. \tag{148}$$

To this end, suppose $I = \{1, \ldots, |I|\}$ wlog and $p_i \equiv |I|^{-1}$ as in (87), and let $\tilde{\ell} \in V_{\mu_{\mathcal{Y}}}$ be any minimizer of $\boldsymbol{\phi}^f_{\mu_{\mathcal{Y}}|\lambda}$. Then $\tilde{\ell} = \sum_{j\in I} \tilde{\alpha}_j q_j$ for some $\tilde{\alpha}^{(\tilde{\ell})} \equiv (\tilde{\alpha}_j) \in \mathbb{R}^{|I|}$ and $q^*_{\tilde{\ell}} \equiv \langle \tilde{\ell}, q\rangle = \sum_{j\in I} \tilde{\alpha}_j q^*_j \in V^*_{\mu_{\mathcal{Y}}}$ minimizes the function $\tilde{\phi}_{f,\lambda}$ from the lemma's last line. Thus, said $\tilde{\alpha}^{(\tilde{\ell})}$ is a global minimizer of

$$\begin{aligned}\mathbb{R}^{|I|} \ni (\alpha_j) \;\longmapsto\; \tilde{F}(\alpha) \coloneqq \tilde{\phi}_{f,\lambda}\big(\textstyle\sum_{j\in I} \alpha_j q^*_j\big) \;&=\; \frac{1}{|I|} \sum_{i=1}^{|I|} \Big| f(y_i) - \sum_{j=1}^{|I|} \alpha_j \langle q_i, q_j\rangle \Big|^2 + \lambda \sum_{i,j=1}^{|I|} \alpha_i \alpha_j \langle q_i, q_j\rangle \\ &=\; \frac{1}{|I|}\Big[\big\|b - \boldsymbol{A}\alpha\big\|_2^2 + |I|\lambda \alpha^{\intercal} \boldsymbol{A} \alpha\Big]\end{aligned} \tag{149}$$

for $\|\cdot\|_2$ the Euclidean norm on $\mathbb{R}^{|I|\times|I|}$ and with the $\mu_{\mathcal{Y}}$-dependent coefficients

$$\boldsymbol{A} \coloneqq \big(\langle q_i, q_j\rangle\big)_{i,j=1}^{|I|} \quad \text{and} \quad b \coloneqq \big(f(y_i)\big)_{i=1}^{|I|}. \tag{150}$$

After expanding and rearranging the terms in (149), we obtain for $F \coloneqq |I|\tilde{F}$ that

$$F(\alpha) = \alpha^{\intercal}(\boldsymbol{A}^{\intercal}\boldsymbol{A} + |I|\lambda \boldsymbol{A})\alpha - 2\alpha^{\intercal}\boldsymbol{A}b \qquad \text{and} \qquad \nabla_\alpha F = 2(\boldsymbol{A}^{\intercal}\boldsymbol{A} + |I|\lambda \boldsymbol{A})\alpha - 2\boldsymbol{A}b.$$

Since the function $F : \mathbb{R}^{|I|} \to \mathbb{R}$ is convex, all of its local minimizers are global minimizers, and so

$$\operatorname*{arg\,min}_{\alpha\in\mathbb{R}^{|I|}} F(\alpha) \;=\; \{\alpha \in \mathbb{R}^{|I|} \mid \nabla_\alpha F = 0\} \;=\; \big\{\alpha \in \mathbb{R}^{|I|} \,\big|\, (\boldsymbol{A}^{\intercal}\boldsymbol{A} + |I|\lambda \boldsymbol{A})\alpha = \boldsymbol{A}b\big\}. \tag{151}$$

Now since $\tilde{\alpha}^{(\tilde{\ell})} \in \arg\min_{\alpha\in\mathbb{R}^{|I|}} F(\alpha)$, the claim (148) thus follows if $\boldsymbol{A}$ (and thus $\boldsymbol{A}^\intercal\boldsymbol{A} + |I|\lambda\boldsymbol{A}$) is invertible. But since $\boldsymbol{A}$ is a Gramian matrix by its definition (150), we know that $\boldsymbol{A}$ is invertible iff

$$\text{the vectors } \; q_1, \dots, q_{|I|} \in \mathcal{H}_\mathcal{Y} \; \text{ are linearly independent.} \tag{152}$$

But under the lemma's premise, namely that the points y_i, $i \in I$, be pairwise distinct, the required independence (152) follows by (Theorem 3.9 and the injectivity of any (48) and) [36, Cor. 2.16].

As to the lemma's final assertion, note that for the (unique) minimizer $h_\star = \arg\min_{h\in\mathcal{H}_{\tilde{\kappa}}} \tilde{\phi}_{f,\lambda}(h)$ we saw that $\mathcal{L}_{h_\star} \coloneqq \{\ell \in \mathcal{H}_\mathcal{Y} \mid \langle \ell, q\rangle = h_\star\} = \{\ell \in \mathcal{H}_\mathcal{Y} \mid \tilde{\phi}_{f,\lambda}(\ell) = \gamma_\star\}$ and hence

$$\|h_\star\|_{\tilde{\kappa}} \overset{\text{def}}{=} \inf\{\|\ell\|_{\mathcal{H}_\mathcal{Y}} \,|\, \ell \in \mathcal{L}_{h_\star}\} \overset{(88)}{=} \big\|\ell^{f,\star}_{\mu_\mathcal{Y},\lambda}\big\|_{\mathcal{H}_\mathcal{Y}}$$

as claimed. □

B.3.3 Proof of Proposition 5.3

Proof. The equivalence between (the first inclusion in) (92) and (93) has been shown in the proof of Lemma 5.2, see the argumentation around (151). The proposition's remaining assertions follow by a similar adaptation of the representer theorem [51]. Indeed: By the definitions of $\boldsymbol{\phi}^{(m)}_{\mu|\lambda}$ and $\hat{\mu}_\mathfrak{Z}$, we have

$$\mathfrak{H}_\Xi^{[m]} \ni g \equiv (g_w)_{w\in[\tilde{d}]^*_{\leq m}} \longmapsto \boldsymbol{\phi}^{(m)}_{\hat{\mu}_\mathfrak{Z}|\lambda}(g) = \frac{1}{N}\sum_{j=1}^N \big\|\pi_{[m]}\big(\tilde{q}(Y^{(j)})\big) - g(X^{(j)})\big\|^2_{\mathcal{H}_\mathcal{Y}} + \lambda\|g\|^2_{\mathfrak{H}_\Xi} = \sum_{k=1}^{m_{\tilde{d}}} \phi_k(g_w) \tag{153}$$

for each $m \in \overline{\mathbb{N}}$, where $\tilde{q} \coloneqq \underline{\mathfrak{sig}}_\Lambda$ and the scalar functions $(\phi_k)_{k\in\mathbb{N}} \subset \mathcal{H}_\kappa$ are defined as

$$\phi_k(h) \coloneqq \frac{1}{N}\sum_{j=1}^N \big|\langle \tilde{q}(Y^{(j)}), \eta^{-1}(k)\rangle - h(X^{(j)})\big|^2 + \lambda\|h\|^2_\kappa.$$

Note that the second identity in (153) is due to (the definition of $\|\cdot\|_{\mathcal{H}_\mathcal{Y}}$ in (42) and) Proposition 5.1, whose first assertion provides the unitary isomorphism $\mathfrak{H}_\Xi \ni g \mapsto (\langle g, w\rangle \mid w \in [\tilde{d}]^*) \in \bigoplus_{w\in[\tilde{d}]^*} \mathcal{H}_\kappa$ (so that $\|(g_w)\|^2_{\mathfrak{H}_\Xi} = \sum_{w\in[\tilde{d}]^*}\langle g_w, g_w\rangle^2_\kappa = \sum_{w\in[\tilde{d}]^*}\|g_w\|^2_\kappa$), cf. also [7, Example 3.2 (iii)]. Therefore,

$$\arg\min_{g\in\mathfrak{H}_\Xi^{[m]}} \boldsymbol{\phi}^{(m)}_{\mu_\mathfrak{Z}|\lambda}(g) \;=\; \arg\min_{(g_w)\in(\mathcal{H}_\kappa)^{\times m_{\tilde{d}}}} \sum_{k=1}^{m_{\tilde{d}}} \phi_k(g_{\eta^{-1}(k)}) \;=\; \left\{ \sum_{k=1}^{m_{\tilde{d}}} g^*_k \cdot \eta^{-1}(k) \;\middle|\; g^*_k \in \arg\min_{h\in\mathcal{H}_\kappa} \phi_k(h) \right\}. \tag{154}$$

Now from the classical representer theorem [51] (cf. also the proof of Lemma 5.2) we know that

$$g^*_k \in \arg\min_{h\in\mathcal{H}_\kappa} \phi_k(h) \quad \text{only if} \quad g^*_k \in \operatorname{span}\big(\kappa(X^{(j)}, \cdot\,) \;\big|\; j \in [N]\big),$$

which together with (154) (and the definition of κ, see Prop. 5.1) proves (91). In fact, as detailed in the proof of Lemma 5.2, for instance, we have that $g^*_k = \arg\min_{h\in\mathcal{H}_\kappa} \phi_k(h)$ if and only if

$$g^*_k = \sum_{j=1}^N \alpha^{(k)}_j \langle q_j, q\rangle \qquad \text{for} \quad \tilde{\alpha}_{(k)} \equiv \big(\alpha^{(k)}_j\big) \in \mathbb{R}^N \quad \text{such that} \quad \boldsymbol{C}_{\mathfrak{Z},\lambda}\tilde{\alpha}_{(k)} = \boldsymbol{A}_\mathfrak{Z} b^{(k)}_\mathfrak{Z}, \tag{155}$$

where $q \coloneqq \underline{\mathfrak{sig}}_\Xi$ and $q_j \coloneqq q(X^{(j)})$ and $\boldsymbol{C}_{\mathfrak{Z},\lambda} \coloneqq (\boldsymbol{A}^\top_\mathfrak{Z}\boldsymbol{A}_\mathfrak{Z} + N\lambda\boldsymbol{A}_\mathfrak{Z})$ for $\boldsymbol{A}_\mathfrak{Z}$ as in (96) and $b^{(k)}_\mathfrak{Z} \coloneqq (\langle \tilde{q}(Y^{(j)}), \eta^{-1}(k)\rangle)_{j\in[N]}$; cf. (151). Now by the definitions (90), (86), (97) and uniqueness (Lem. 5.2),

$$\hat{\Upsilon}^{\text{lex}}_{N,\lambda,m} = \sum_{k=1}^{m_{\tilde{d}}} g^*_k \cdot e_k \overset{(155)}{=} \sum_{k=1}^{m_{\tilde{d}}} \left[\sum_{j=1}^N \alpha^{(k)}_j q^*_j\right] \cdot e_k = \Big(\tilde{\alpha}^\intercal_{(1)} \cdot \underline{q}^* \Big| \cdots \Big| \tilde{\alpha}^\intercal_{(m_{\tilde{d}})} \cdot \underline{q}^*\Big)^\intercal = \tilde{A}^\intercal_* \cdot \underline{q}^*$$

for the matrix $\tilde{A}_* \equiv \big(\tilde{\alpha}_{(1)}|\cdots|\tilde{\alpha}_{(m_{\tilde{d}})}\big) \in \mathbb{R}^{N\times m_{\tilde{d}}}$ and the function $\underline{q}^* \coloneqq \sum_{j=1}^N q_j^* \cdot \tilde{e}_j : \mathcal{X} \to \mathbb{R}^N$, where $(e_k)_{k\in[m_{\tilde{d}}]}$ and $(e_j)_{j\in[N]}$ are the standard bases of $\mathbb{R}^{m_{\tilde{d}}}$ and $\mathbb{R}^N$, respectively; this completes the proof of (92). Noting that $\boldsymbol{B}_{\mathfrak{Z}} = \big(b_{\mathfrak{Z}}^{(1)}|\cdots|b_{\mathfrak{Z}}^{(m_{\tilde{d}})}\big)$, equation (94) is clear from (155). □

B.3.4 Proof of Lemma 5.5

Proof. The statements (i) and (iii) both follow from essentially the same simple observations: By the fact that $\mathcal{H}_\kappa$ is dense in $L^2(\mathbb{P}_Y)$ (Proposition 5.1) and from the definition (84), we for each $\epsilon > 0$ with $\vartheta_f(\epsilon) \neq \infty$ can find $\ell_\epsilon \in \mathcal{H}_\mathcal{Y}$ such that $\boldsymbol{\phi}^f_{\mathbb{P}_Y}(\ell_\epsilon) \leq \epsilon^2/2$ and $\|\langle \ell_\epsilon, \underline{\mathfrak{sig}}_\Lambda\rangle\|_\kappa^2 \leq \epsilon^2/2\vartheta_f(\epsilon)$. Hence for each $\lambda > 0$, we have by definition of ℓ^f_λ and the property $\boldsymbol{\phi}^f_{\mathbb{P}_Y}(\ell_\epsilon) \leq \epsilon^2/2$ that

$$\begin{aligned} r^{\mathrm{I}}_f(\lambda)^2 = \boldsymbol{\phi}^f_{\mathbb{P}_Y}(\ell^f_\lambda) \;&\leq\; \boldsymbol{\phi}^f_{\mathbb{P}_Y|\lambda}(\ell^f_\lambda) = \boldsymbol{\phi}^f_{\mathbb{P}_Y|\lambda}(\ell^f_\lambda) - \boldsymbol{\phi}^f_{\mathbb{P}_Y}(\ell_\epsilon) + \boldsymbol{\phi}^f_{\mathbb{P}_Y}(\ell_\epsilon) \\ &\leq\; \boldsymbol{\phi}^f_{\mathbb{P}_Y|\lambda}(\ell_\epsilon) - \boldsymbol{\phi}^f_{\mathbb{P}_Y}(\ell_\epsilon) + \boldsymbol{\phi}^f_{\mathbb{P}_Y}(\ell_\epsilon) \;\leq\; \lambda\big\|\langle \ell_\epsilon, \underline{\mathfrak{sig}}_\Lambda\rangle\big\|_\kappa^2 + \frac{\epsilon^2}{2}, \end{aligned} \tag{156}$$

which (by the above choice of ℓ_ϵ) implies (i) for the case $\vartheta_f(\epsilon) \neq \infty$. If $\vartheta_f(\epsilon) = \infty$, then for each $\lambda > 0$ there will be some $\ell^\lambda_\epsilon \in \mathcal{H}_\mathcal{Y}$ with $\boldsymbol{\phi}^f_{\mathbb{P}_Y}(\ell^\lambda_\epsilon) \leq \epsilon^2/2$ and $\|\langle \ell_\epsilon, \underline{\mathfrak{sig}}_\Lambda\rangle\|_\kappa^2 \leq \epsilon^2/2\lambda$, which, by the fact that the inequalities (156) are valid for all $(\lambda, \ell_\epsilon) \in \mathbb{R}_{>0} \times \big(\boldsymbol{\phi}^f_{\mathbb{P}_Y}\big)^{-1}(0, \epsilon^2/2]$, implies (i) also for the case $\vartheta_f(\epsilon) = \infty$. Statement (iii) follows analogously to (i) upon noting that, by statement (ii),

$$\begin{aligned} R^{\mathrm{I}}_m(\lambda)^2 = \;& \boldsymbol{\phi}^{(m)}_{\mathbb{P}_{(X,Y)}}\big(\pi_{[m]}(\Upsilon^*_\lambda)\big) - \boldsymbol{\phi}^{(m)}_{\mathbb{P}_{(X,Y)}}\big(\pi_{[m]}(\varphi_\star)\big) \\ \leq\; & \boldsymbol{\phi}^{(m)}_{\mathbb{P}_{(X,Y)}|\lambda}(g) - \boldsymbol{\phi}^{(m)}_{\mathbb{P}_{(X,Y)}}(g) + \Delta_g \;\leq\; \lambda\|g\|^2_{\mathfrak{H}_\Xi} + \Delta_g \end{aligned} \tag{157}$$

for the difference $\Delta_g \coloneqq \boldsymbol{\phi}^{(m)}_{\mathbb{P}_{(X,Y)}}(g) - \boldsymbol{\phi}^{(m)}_{\mathbb{P}_{(X,Y)}}\big(\pi_{[m]}(\varphi_\star)\big)$ and any $\lambda > 0$ and any $g \in L^2(\mathbb{P}_X; \mathcal{H}^{[m]}_\mathcal{Y})$; then exactly as shown for (i) above, the density result [following (81)] of Proposition 5.1 allows us to, for any given $\epsilon > 0$ and any $0 < \lambda \leq \tilde{\vartheta}_{\pi_{[m]}(\varphi_\star)}(\epsilon)$, find a $g \in L^2(\mathbb{P}_X; \mathcal{H}^{[m]}_\mathcal{Y})$ such that $\lambda\|g\|^2_{\mathfrak{H}_\Xi}$ and Δ_g are both bounded by $\epsilon^2/2$ (the case $\tilde{\vartheta}_{\pi_{[m]}(\varphi_\star)}(\epsilon) = \infty$ is handled as before), implying (iii).

For the proof of (157), that is of statement (ii), abbreviate $\varphi^{[m]}_\star \coloneqq \pi_{[m]}(\varphi_\star)$ and note that with

$$\varphi^{[m]}_\star(X) = \pi_{[m]}\big(\mathbb{E}[\Upsilon^\Lambda \,|\, X]\big) = \mathbb{E}[\pi_{[m]}(\Upsilon^\Lambda) \,|\, X] = \varphi^\star_m(X) \quad \text{for} \quad \varphi^\star_m \coloneqq \operatorname*{arg\,min}_{g \,\in\, L^2(\mathbb{P}_X; \mathcal{H}^{[m]}_\mathcal{Y})} \boldsymbol{\phi}^{(m)}_{\mathbb{P}_{(X,Y)}}(g)$$

(where the second of the above equalities holds by the commuting property (75) and the linearity of conditional expectations, and the third equality holds by the variational characterisation (cf. (141)) of conditional expectations) we have $\varphi^{[m]}_\star = \varphi^\star_m$ in $L^2(\mathbb{P}_X; \mathcal{H}^{[m]}_\mathcal{Y})$ and hence, as claimed, obtain that

$$R^{\mathrm{I}}_m(\lambda)^2 = \mathbb{E}\Big[\big\|\pi_{[m]}(\Upsilon^*_\lambda) - \varphi^\star_m(X)\big\|^2_{\mathcal{H}_\mathcal{Y}}\Big] = \boldsymbol{\phi}^{(m)}_{\mathbb{P}_{(X,Y)}}\big(\pi_{[m]}(\Upsilon^*_\lambda)\big) - \boldsymbol{\phi}^{(m)}_{\mathbb{P}_{(X,Y)}}\big(\varphi^{[m]}_\star\big)$$

for each $\lambda > 0$, by the classical risk decomposition from [16, Proposition 1].

As to statement (iv), denote $\tilde{q}_l \coloneqq \underline{\mathfrak{sig}}_\Lambda(Y^{(l)})$ and note that, by the definitions (98) and (42),

$$\big\|\pi_\nu(\hat{\ell}^\star_{n,\lambda,f})\big\|_\nu \;\leq\; \sum_{l=1}^n \big|\alpha^\star_l\big|\|\pi_\nu(\tilde{q}_l)\|_\nu \;\leq\; \|\alpha^\star\|_2\gamma_\nu \quad \text{for} \quad \gamma_\nu \coloneqq \sqrt{\sum_{l=1}^n \|\pi_\nu(\tilde{q}_l)\|^2_\nu} \qquad \big(\nu \in \mathbb{N}\big).$$

Hence by the definitions of $r^{\mathrm{III}}_{f,\lambda,n}(m)$ and $\pi_{[m]}$ and (again) with definition (42), for each $m \in \mathbb{N}_0$,

$$r^{\mathrm{III}}_{f,\lambda,n}(m) \;=\; \sqrt{\sum_{\nu=m+1}^\infty \big\|\pi_\nu(\hat{\ell}^\star_{n,\lambda,f})\big\|^2_\nu} \;\leq\; \|\alpha^\star\|_2\tilde{\beta}_m =: \beta^{(\mathrm{cut})}_m \qquad \text{for} \quad \tilde{\beta}_m \coloneqq \sqrt{\sum_{\nu=m+1}^\infty \gamma^2_\nu}. \tag{158}$$

Now since $\tilde{q}_1, \dots, \tilde{q}_n \in \mathcal{H}_\mathcal{Y}$ and thus $\sum_{\nu=0}^{\infty} \gamma_\nu^2 = \sum_{l=1}^{n} \|\tilde{q}_l\|_{\mathcal{H}_\mathcal{Y}} < \infty$, we find that $\big(\beta_m^{(\mathrm{cut})}\big)_{m=1}^{\infty}$ is a monotone decreasing null sequence which, due to $\alpha^\star = (\boldsymbol{A}_\mathfrak{Y}^\top \boldsymbol{A}_\mathfrak{Y} + n\lambda \boldsymbol{A}_\mathfrak{Y})^+ \boldsymbol{A}_\mathfrak{Y} b_{\mathfrak{Y},f}$, depends on $(\mathfrak{Y}, \Lambda, \lambda, f)$ only.

For (v), let $\delta > 0$ and observe that by [42, Proposition 3.2], (102) then holds for the sequences

$$\beta_n^{(\mathrm{out})} \coloneqq \frac{c_\Lambda}{\lambda\sqrt{n\delta}} \qquad \text{and} \qquad \beta_N^{(\mathrm{in})} \coloneqq \sqrt{\frac{c_\Xi c_\Lambda}{\lambda^2 N \delta}} \qquad (n, N \in \mathbb{N})$$

featuring the constants $c_Z \coloneqq \sup_{\boldsymbol{t} \in \mathcal{H}_\xi} \|Z(\boldsymbol{t})\|_{\mathcal{H}_\xi}^2 < \infty$ for $(Z, \xi) \in \{(\Lambda, \mathcal{Y}), (\Xi, \mathcal{X})\}$ (recall Definition 3.18 for their finiteness); recall that the general applicability of [42] is due to Proposition 5.1. □

B.3.5 Algorithms I and II are Instances of Definition 2.6

Remark B.7. An approximation of the regression operator (14), the estimator (107) can be written

$$\hat{T}_N^{\mathrm{I}}(f, \mathfrak{Y}; \mathfrak{Z}, \underline{\lambda}, n, m) \coloneqq \hat{\mathscr{R}}_{\underline{\lambda}, f}^{\mathrm{I}}[n, N, m] = \Big\langle \hat{\boldsymbol{\ell}}_{\lambda_1;\mathfrak{Y}}^f \,,\, \hat{\boldsymbol{\Xi}}_{\lambda_2,n,m;\mathfrak{Z}}(\cdot) \Big\rangle_{2,k} \tag{159}$$

—that is: in the form of Definition 2.6, eq. (18), and (24)—for the bilinear map

$$\big\langle \cdot, \cdot \big\rangle_{2,k} : \ell^2(\mathbb{N}_0)^{\times k} \times \ell^2(\mathbb{N}) \longrightarrow \mathbb{R}^k, \ \big((\mathfrak{q}_i), \mathfrak{q}\big) \mapsto \big(\langle \mathfrak{q}_i, \mathfrak{q} \rangle_{\ell^2}\big)_{i=1,\dots,k}$$

and the constituent component representations

$$\hat{\boldsymbol{\ell}}_{\lambda_1;\mathfrak{Y}}^f \coloneqq \Big(\sum_{l=1}^{n} \big[\boldsymbol{\alpha}_{\lambda_1}^f\big]_{li} \eta\big(\underline{\mathfrak{sig}}_\Lambda(Y^{(l)})\big)\Big)_{i=1,\dots,k} \quad \text{and} \quad \hat{\boldsymbol{\Xi}}_{\lambda_2,n,m;\mathfrak{Z}} \coloneqq \hat{\Upsilon}_{N,\lambda_2,m}^{\mathrm{lex}} : \mathcal{X} \to \mathbb{R}^{m_{\tilde{d}}},$$

where $\eta : \mathcal{H}_\mathcal{Y} \to \ell^2(\mathbb{N}_0)$, $\big(\boldsymbol{t}_w\big)_{w \in [\tilde{d}]^*} \mapsto \big(\boldsymbol{t}_{\eta^{-1}(j)}\big)_{j \in \mathbb{N}_0}$ extends the shortlex ordering used in (97). Note that the estimator (159) is linear in both $\hat{\boldsymbol{\ell}}_{\lambda_1;\mathfrak{Y}}^f \in \ell^2(\mathbb{N}_0)^{\times k}$ and $f \in \mathcal{L}^2(\mathbb{P}_Y)$. ◆

Remark B.8. An approximation of the regression operator (14), the estimator (115) can be written

$$\hat{T}_M^{\mathrm{II}}(f; \mathfrak{W}_f, \lambda) \coloneqq \hat{\mathscr{R}}_{\lambda, f(Y)}^{\mathrm{II}}[M] = \Big\langle \hat{\boldsymbol{\varrho}}_{\lambda;\mathfrak{W}_f}, \underline{\mathfrak{sig}}_\Xi(\cdot) \Big\rangle_{\mathcal{H}_\mathcal{X},\, k} \tag{160}$$

for the data $\mathfrak{W}_f \coloneqq \{(X^{(j)}, f(Y^{(j)})) \mid j \in [M]\}$, which are iid copies of $(X, f(Y))$, the bilinear map

$$\big\langle \cdot, \cdot \big\rangle_{\mathcal{H}_\mathcal{X},\, k} : \mathcal{H}_\mathcal{X}^{\times k} \times \mathcal{H}_\mathcal{X} \longrightarrow \mathbb{R}^k, \ \big((\mathfrak{q}_i), \mathfrak{q}\big) \mapsto \big(\langle \mathfrak{q}_i, \mathfrak{q} \rangle_{\mathcal{H}_\mathcal{X}}\big)_{i=1,\dots,k},$$

and the constituent $\mathfrak{W}_f$-processing data representation, with $\hat{A}_* = (\hat{\alpha}_{ij}^{(\lambda)})$ solving $(114)|_{Z=f(Y)}$,

$$\hat{\boldsymbol{\varrho}}_{\lambda;\mathfrak{W}_f} \coloneqq \Big(\sum_{j=1}^{M} \big[\hat{A}_*\big]_{ij} \underline{\mathfrak{sig}}_\Xi(X^{(j)})\Big)_{i=1,\dots,k}.$$

Note that while the estimator (160) is also linear[11] in $f \in \mathcal{L}^2(\mathbb{P}_Y)$ and $\hat{\boldsymbol{\varrho}}_{\lambda;\mathfrak{W}_f} \in \mathcal{H}_\mathcal{X}^{\times k}$, unlike the more modular estimator (160) it does not us a factorisation into two separate and possibly independent types of data (i.e., $\mathfrak{Z}$ and $\mathfrak{Y}_f \coloneqq \{(Y_l, f(Y_l)) \mid Y_l \in \mathfrak{Y}\}$) but instead operates directly on the combined dataset $\mathfrak{W}_f$, where $f(Y)$ is observed exclusively in association with X. ◆

[11] In the sense that: $\hat{T}_M^{\mathrm{II}}(f_1 + cf_2; \mathfrak{W}_{f_1+cf_2}, \lambda) = \hat{T}_M^{\mathrm{II}}(f_1; \mathfrak{W}_{f_1}, \lambda) + c\hat{T}_M^{\mathrm{II}}(f_2; \mathfrak{W}_{f_2}, \lambda)$ for any $f_1, f_2 \in \mathcal{L}^2(\mathbb{P}_Y), c \in \mathbb{R}$.

B.3.6 Proof of Theorem 5.9

Proof. Denote $\psi \coloneqq \mu_{\cdot}(f)$, so that $\psi \equiv (\psi_1, \cdots, \psi_k) \in L^2(\mathbb{P}_X; \mathbb{R}^k)$. Showing $(121)|_{\nu=\mathrm{I}}$ first, fix any $(\varepsilon, \delta) \in \mathbb{R}^2_{>0}$ and take an arbitrary $(\underline{\lambda}, n, N, m) \in \mathcal{A}^{\mathrm{I}}_f(\varepsilon, \delta)$. Recalling from (159) that

$$\hat{\mathscr{R}}^{\mathrm{I}}_{\underline{\lambda},f}[n, N, m] = \sum_{i=1}^{k} \langle \hat{\ell}^i_{\lambda_1,n,m}, \hat{\Upsilon}^*_{N,\lambda_2,m} \rangle e_i \tag{161}$$

with $\hat{\Upsilon}^*_{N,\lambda_2,m}$ as in (90) and with $\hat{\ell}^i_{\lambda_1,n,m} \coloneqq \pi_{[m]}(\hat{\ell}^\star_{n,\lambda_1,f_i})$ for $\hat{\ell}^\star_{n,\lambda_1,f_i}$ as in (98), we find that

$$\mathbb{P}\big(\big\|\psi - \hat{\mathscr{R}}^{\mathrm{I}}_{f,\underline{\lambda}}[n, N, m]\big\| \geq \varepsilon\big) \;\leq\; \mathbb{P}(E_1 \geq 3\varepsilon/5) + \mathbb{P}(E_2 \geq 2\varepsilon/5) \tag{162}$$

(from (161) and by the triangle inequality) with $\|\cdot\| \coloneqq \|\cdot\|_{L^2(\mathbb{P}_X;\mathbb{R}^k)}$ and for the differences

$$E_1 \coloneqq \big\|\psi - \hat{\psi}_{\lambda_1,n,m}\big\| \qquad \text{and} \qquad E_2 \coloneqq \Big\|\textstyle\sum_{i=1}^k \langle \hat{\ell}^i_{\lambda_1,n,m}, \varphi_\star - \hat{\Upsilon}^*_{N,\lambda_2,m} \rangle e_i\Big\|,$$

where $\hat{\psi}_{\lambda_1,n,m} \coloneqq \sum_{i=1}^k \langle \hat{\ell}^i_{\lambda_1,n,m}, \varphi_\star \rangle e_i$ and for $\varphi_\star$ as in (101). Using (75) and the inequality in (74),

$$E_1 \;\leq\; \sum_{i=1}^{k} \tilde{E}_{1,i} \qquad \text{for} \quad \tilde{E}_{1,i} \coloneqq \big\|f_i - \langle \hat{\ell}^i_{\lambda_1,n,m}, \underline{\mathfrak{sig}}_\Lambda \rangle\big\|_{L^2(\mathbb{P}_Y)}.$$

Recalling the definitions in (99) and (100), we obtain, for each $i \in [k]$, that

$$\tilde{E}_{1,i} \;\leq\; r^{\mathrm{I}}_{f_i}(\lambda_1) + r^{\mathrm{II}}_{f_i,\lambda_1}(n) + r^{\mathrm{III}}_{f_i,\lambda_1,n}(m) \;\leq\; \frac{\varepsilon}{5k} + r^{\mathrm{II}}_{f_i,\lambda_1}(n) + \frac{\varepsilon}{5k}\,,$$

where we used the error bound $\max\big\{r^{\mathrm{I}}_{f_i}(\lambda_1), r^{\mathrm{III}}_{f_i,\lambda_1,n}(m)\big\} \leq \varepsilon/(5k)$ which is due to Lemma 5.5 (i) and (iv) in conjunction with the $\mathcal{B}_f(\varepsilon,\delta)$-defining constraints defined in (104) and (105). Hence,

$$\mathbb{P}(E_1 \geq 3\varepsilon/5) \;\leq\; \mathbb{P}\left(\sum_{i=1}^{k} r^{\mathrm{II}}_{f_i,\lambda_1}(n) \geq \varepsilon/5\right) \;\leq\; \sum_{i=1}^{k} \mathbb{P}\Big(r^{\mathrm{II}}_{f_i,\lambda_1}(n) \geq \varepsilon/(5k)\Big) \;\leq\; \frac{\delta}{2}\,, \tag{163}$$

where the last inequality is due to Lemma 5.5 (v) via the constraint $n \geq n_{f,\lambda_1}(\varepsilon,\delta)$.

For a bound on $\mathbb{P}(E_2 \geq 2\varepsilon/5)$, note that, by Cauchy-Schwarz and the triangle inequality,

$$E_2 \;\leq\; c_{\lambda_1,m;\mathfrak{Y}} \big\|\pi_{[m]}(\varphi_\star) - \hat{\Upsilon}^*_{N,\lambda_2,m}\big\|_{L^2(\mathbb{P}_X;\mathcal{H}_\mathcal{Y})} \;\leq\; c_{\lambda_1,m;\mathfrak{Y}} \big(R^{\mathrm{I}}_m(\lambda_2) + R^{\mathrm{II}}_{m,\lambda_2}(N)\big) \tag{164}$$

with $c_{\lambda_1,m;\mathfrak{Y}}$ as defined in (106) and for the errors R^{I}_m and $R^{\mathrm{II}}_{m,\lambda}$ introduced in (101). Lemma 5.5 (iii) and the bound $\lambda_2 \leq \lambda^{\mathrm{II}}_{\lambda_1,m}(\varepsilon;\mathfrak{Y})$ (cf. (106)) ensure $c_{\lambda_1,m;\mathfrak{Y}} R^{\mathrm{I}}_m(\lambda_2) \leq \varepsilon/5$. Consequently,

$$\mathbb{P}(E_2 \geq 2\varepsilon/5) \;\leq\; \mathbb{P}\big(R^{\mathrm{II}}_{m,\lambda_2}(N) \geq \varepsilon/(5c_{\lambda_1,m;\mathfrak{Y}})\big) \;\leq\; \frac{\delta}{2} \tag{165}$$

follows by combining (164) and Lemma 5.5 (v), using that $N \geq N_{\lambda_1,\lambda_2,m}(\varepsilon,\delta;\mathfrak{Y})$ (cf. (106)).

Applying the inequalities (163) and (165) to (162) proves the case $\nu = \mathrm{I}$ in (121), as desired.

The proof of $(121)|_{\nu=\mathrm{II}}$ is very similar but more concise, using Corollary 5.8 applied to $Z \coloneqq f(Y)$ instead of Lemma 5.5: For any fixed $(\lambda, N) \in \mathcal{A}_f(\varepsilon,\delta)$, we know that (since $\psi = \psi_{X,Z}$ by uniqueness)

$$\mathbb{P}\big(\big\|\psi - \hat{\Upsilon}^*_{f(Y),\lambda;N}\big\| \geq \varepsilon\big) \;\leq\; \mathbb{P}(\rho^{\mathrm{I}}_Z(\lambda) \geq \varepsilon/2) + \mathbb{P}(\rho^{\mathrm{II}}_{Z,\lambda}(N) \geq \varepsilon/2) \;\leq\; 0 + \delta$$

by (the triangle inequality and) Corollary 5.8 (ii) and (iii); this proves the case $\nu = \mathrm{II}$ in (121).

Proving (122), recall that $\hat{\mathscr{R}}^{\mathrm{I}}_{\theta} \coloneqq \hat{\mathscr{R}}^{\mathrm{I}}_{f,\underline{\lambda}}[n,N,m]$ for $\theta = (\underline{\lambda},n,N,m)$, and $\hat{\mathscr{R}}^{\mathrm{II}}_{\theta} \coloneqq \hat{\mathscr{R}}^{\mathrm{II}}_{\lambda,Z}[M]$ for $\theta = (\lambda,N)$. Then for both $\nu = \mathrm{I},\mathrm{II}$, the object $\hat{\mathscr{R}}^{\nu}_{\theta}$ is an $L^2(\mathbb{P}_X;\mathbb{R}^k)$-valued random variable:

$$\hat{\mathscr{R}}^{\nu}_{\theta} = \hat{\mathscr{R}}_{\theta}(\,\cdot\,;\tilde{Z}_{\nu}) : \Omega \longrightarrow L^2(\mathbb{P}_X;\mathbb{R}^k),\ \omega \mapsto \hat{\mathscr{R}}_{\theta}\big(\,\cdot\,;\tilde{Z}_{\nu}(\omega)\big),$$

for $\tilde{Z}_{\mathrm{I}}$ resp. $\tilde{Z}_{\mathrm{II}}$ the random vectors from which the data in (89) resp. (112) is sampled, that is

$$\tilde{Z}_{\mathrm{I}} \coloneqq \big(Y^{(1)},\cdots,Y^{(n)},(X^{(1)},Y^{(1)}),\cdots,(X^{(N)},Y^{(N)})\big) : \Omega \longrightarrow \tilde{\mathcal{Z}}^{\mathrm{I}} \coloneqq \mathcal{Y}^n \times \big(\mathcal{X}\times\mathcal{Y}\big)^N$$
$$\text{and} \qquad \tilde{Z}_{\mathrm{II}} \coloneqq \big((X^{(1)},Z_{(1)}),\cdots,(X^{(M)},Z_{(M)})\big) : \Omega \longrightarrow \tilde{\mathcal{Z}}^{\mathrm{II}} \coloneqq \big(\mathcal{X}\times\mathbb{R}^k\big)^M.$$

Define $A^{\nu}_{\theta,\varepsilon} \coloneqq \big\{|\mathbb{E}[f(Y)\,|\,X] - \hat{\mathscr{R}}^{\nu}_{\theta}(X)| \geq \varepsilon\big\} \equiv \big\{\omega\in\Omega\,\big|\,|\mathbb{E}[f(Y)\,|\,X](\omega) - \hat{\mathscr{R}}^{\nu}_{\theta}(X(\omega);\tilde{Z}_{\nu}(\omega))| \geq \varepsilon\big\}$, $A^{\times|\nu}_{\theta,\varepsilon} \coloneqq \big\{(x,z)\in\mathcal{X}\times\tilde{\mathcal{Z}}^{\nu}\,\big|\,|\psi(x) - \hat{\mathscr{R}}^{\nu}_{\theta}(x;z)| \geq \varepsilon\big\}$ and $\Delta^{z}_{\theta|\nu} \coloneqq \big|\psi - \hat{\mathscr{R}}^{\nu}_{\theta}(\,\cdot\,;z)\big|$ for any fixed $z\in\tilde{\mathcal{Z}}^{\nu}$. With the z-sections $A^{z|\nu}_{\theta,\varepsilon} \coloneqq \{x\in\mathcal{X}\mid(x,z)\in A^{\times}_{\theta,\varepsilon}\}$ $(z\in\tilde{\mathcal{Z}}^{\nu})$ for $\nu=\mathrm{I},\mathrm{II}$, we have by that

$$\mathbb{P}(A^{\nu}_{\theta,\varepsilon}) \;=\; \int_{\mathcal{X}\times\tilde{\mathcal{Z}}_{\nu}} \mathbb{1}_{A^{\times|\nu}_{\theta,\varepsilon}}(x,z)\,\mathbb{P}_{(X,\tilde{Z}_{\nu})}(\mathrm{d}x,\mathrm{d}z) \;=\; \int_{\tilde{\mathcal{Z}}_{\nu}} \mathbb{P}_X(A^{z|\nu}_{\theta,\varepsilon})\,\mathbb{P}_{\tilde{Z}_{\nu}}(\mathrm{d}z), \tag{166}$$

which holds by Fubini and the (89)- resp. (112)-underlying assumption that the random variables X and $\tilde{Z}_{\mathrm{I}}$ and $\tilde{Z}_{\mathrm{II}}$ are statistically independent. Now for any $(\tilde{\varepsilon},\tilde{\delta})\in\mathbb{R}^2_{>0}$, the inequalities (121) yield exceptional events $\mathcal{C}^{\nu}_{\tilde{\varepsilon},\tilde{\delta}}\in\mathcal{B}(\tilde{\mathcal{Z}}^{\nu})$ of measure $\mathbb{P}_{\tilde{Z}_{\nu}}(\mathcal{C}^{\nu}_{\tilde{\varepsilon},\tilde{\delta}}) \leq \tilde{\delta}$ such that, for each $\theta\in\mathcal{A}^{\nu}_{f}(\tilde{\varepsilon},\tilde{\delta})$,

$$\mathbb{P}_X(A^{z|\nu}_{\theta,\varepsilon}) \leq \varepsilon^{-2}\big\|\mathbb{1}_{A^{z|\nu}_{\theta,\varepsilon}}\cdot\Delta^{z}_{\theta|\nu}\big\|^2_{L^2(\mathbb{P}_X;\mathbb{R}^k)} \leq \varepsilon^{-2}\|\Delta^{z}_{\theta|\nu}\|^2_{L^2(\mathbb{P}_X;\mathbb{R}^k)} < \frac{\tilde{\varepsilon}}{\varepsilon^2}, \quad \text{for all} \quad z\in\big(\mathcal{C}_{\tilde{\varepsilon},\tilde{\delta}}\big)^{\mathrm{c}}; \tag{167}$$

this holds for both $\nu=\mathrm{I},\mathrm{II}$. The combination of (167) with (166) thus yields that

$$\mathbb{P}(A^{\nu}_{\theta,\varepsilon}) \;\leq\; \int_{\big(\mathcal{C}_{\tilde{\varepsilon},\tilde{\delta}}\big)^{\mathrm{c}}} \mathbb{P}_X(A^{z|\nu}_{\theta,\varepsilon})\,\mathbb{P}_{\tilde{Z}_{\nu}}(\mathrm{d}z) + \tilde{\delta} \;<\; \frac{\tilde{\varepsilon}}{\varepsilon^2}(1-\tilde{\delta}) + \tilde{\delta} \;\leq\; (1-q) =: q' \qquad (\nu=\mathrm{I},\mathrm{II})$$

for each $\theta\in\Theta^{\nu}_{f}(\varepsilon,\delta) \equiv \bigcup_{(\tilde{\varepsilon},\tilde{\delta})\,:\,\tilde{\varepsilon}(1-\tilde{\delta})+\tilde{\delta}\varepsilon^2\,\leq\,q'\varepsilon^2}\mathcal{A}^{\nu}_{f}(\tilde{\varepsilon},\tilde{\delta})$. Since $\mathbb{P}\Big(\big|\mathbb{E}[f(Y)\,|\,X] - \hat{\mathscr{R}}^{\nu}_{\theta}(X)\big| < \varepsilon\Big) = 1 - \mathbb{P}(A^{\nu}_{\theta,\varepsilon})$ for $\nu=\mathrm{I},\mathrm{II}$, the assertion (122) follows. □

B.4 'Stabilized' Estimation of Conditional Expectations (Over L^2-Balls)

To allow for a statistically consistent sample-based approximation of the optimisation scheme (66), we adopt the proof of Theorem 4.8 for a 'stabilized' version of (65).

Lemma B.9. *Given any $R\geq 0$, consider the set defined by*

$$\mathfrak{C}_R \coloneqq \Big\{\,\mathbb{Z}\in L^2_X(\mathcal{H}_{\mathcal{Y}})\;\Big|\;\|\mathbb{Z}\|_{\mathcal{L}^\infty} \coloneqq \operatorname{ess\,sup}_{\omega\in\Omega}\|\mathbb{Z}(\omega)\| \leq R\,\Big\}.$$

Then the following hold.

(i) *The set $\mathfrak{C}_R$ is a convex and complete subset of $L^2_X(\mathcal{H}_{\mathcal{Y}})$.*

(ii) *Denoting $\Psi_R \coloneqq \big\{\psi\cdot\mathbb{1}_{\{\|\psi\|\leq R\}}\,\big|\,\psi\in\Psi\big\}$ (cf. (58)) for $R\geq 0$, then each element of $\mathfrak{C}_R$ is an $\|\cdot\|_{L^2(\mathcal{H}_{\mathcal{Y}})}$-limit point of the set $\Psi_{2R}(X) \coloneqq \{\tilde{\psi}(X)\mid\tilde{\psi}\in\Psi_{2R}\}\subset L^2_X(\mathcal{H}_{\mathcal{Y}})$.*

(iii) *For any $R\geq 0$ and with $\mathfrak{D}_X$ the spatial support of X (see* Rem. B.3 (iii)*), we have that*

$$\mathfrak{C}_R = \big\{\phi(X)\,\big|\,\phi\in\mathcal{B}_R\big\} \quad \textit{with} \quad \mathcal{B}_R \coloneqq \{\phi:\mathfrak{D}_X\to\mathcal{H}_{\mathcal{Y}}\mid\phi \textit{ Borel-measurable}:\ \|\phi\|_{\infty;\mathfrak{D}_X}\leq R\},$$

which holds as an identity in $2^{L^2(\mathcal{H}_{\mathcal{Y}})}$, that is up to element-wise inequality on a $\mathbb{P}$-nullset.

(iv) *We have that* $\mathbb{E}\big[\mathbb{Y}^\Lambda\big|X\big] \in \mathfrak{C}_{R_\Lambda}$ *for* $R_\Lambda \coloneqq \sup\{\|\Lambda(\boldsymbol{t})\| \mid \boldsymbol{t} \in \mathcal{H}_\mathcal{Y}\} < \infty$.

Proof. Fix any $R \geq 0$. The convexity of $\mathfrak{C}_R$ is clear since $\|\cdot\|_{\mathcal{L}^\infty}$ is homogeneous and subadditive.

To see that $\mathfrak{C}_R$ is complete, let $(\mathbb{Z}_k) \subset \mathfrak{C}_R$ be a Cauchy sequence in $(\mathfrak{C}_R, \|\cdot\|_{L^2(\mathcal{H}_\mathcal{Y})})$. There is then $\mathbb{Z} \in L^2_X(\mathcal{H}_\mathcal{Y})$ with $\lim_{k\to\infty}\|\mathbb{Z}-\mathbb{Z}_k\|_{L^2(\mathcal{H}_\mathcal{Y})} = 0$, by the fact that $L^2_X(\mathcal{H}_\mathcal{Y})$ is complete. To see that in fact $\mathbb{Z} \in \mathfrak{C}_R$, note that for each $\varepsilon > 0$ and with $A_\varepsilon \coloneqq \{\|\mathbb{Z}\| - R \geq \varepsilon\}$,

$$\mathbb{P}(A_\varepsilon) \;=\; \frac{1}{\varepsilon^2}\int_{A_\varepsilon} \varepsilon^2 \,\mathrm{d}\mathbb{P} \;\leq\; \frac{1}{\varepsilon^2}\int_{A_\varepsilon} \|\mathbb{Z}-\mathbb{Z}_k\|^2 \,\mathrm{d}\mathbb{P} \;\leq\; \frac{1}{\varepsilon^2}\|\mathbb{Z}-\mathbb{Z}_k\|^2_{L^2(\mathcal{H}_\mathcal{Y})} \quad \text{for all} \;\; k \in \mathbb{N},$$

hence $\mathbb{P}(A_\varepsilon) = 0$ and thus, as $\varepsilon > 0$ was arbitrary, $\mathbb{P}(\|Z\| > R) \leq \mathbb{P}(\bigcup_{n\in\mathbb{N}} A_{1/n}) \leq \sum_{n\in\mathbb{N}} \mathbb{P}(A_{1/n}) = 0$. This implies that $\|\mathbb{Z}\|_{\mathcal{L}^\infty} \leq R$ and hence $\mathbb{Z} \in \mathfrak{C}_R$, concluding the proof of (i).

For a proof of statement (ii), choose an arbitrary $\mathbb{Z}_\star \in \mathfrak{C}_R$ and any $\varepsilon > 0$. Since by Proposition 4.5 there is $\psi \in \Psi$ with $\|\mathbb{Z}_\star - \psi(X)\|_{L^2(\mathcal{H}_\mathcal{Y})} \leq \varepsilon$, the choice $\tilde{\psi} \coloneqq \psi \cdot \mathbb{1}_{\{\|\psi\|\leq 2R\}}$ gives that

$$\begin{aligned}
\big\|\mathbb{Z}_\star - \tilde{\psi}(X)\big\|^2_{L^2(\mathcal{H}_\mathcal{Y})} \;&=\; \int_{\mathcal{A}_R} \big\|\mathbb{Z}_\star - \psi(X)\big\|^2 \,\mathrm{d}\mathbb{P} + \int_{\mathcal{A}_R^c} \|\mathbb{Z}_\star\|^2 \,\mathrm{d}\mathbb{P} \\
&\leq\; \int_{\mathcal{A}_R} \big\|\mathbb{Z}_\star - \psi(X)\big\|^2 \,\mathrm{d}\mathbb{P} + \int_{\mathcal{A}_R^c} \|\mathbb{Z}_\star - \psi(X)\|^2 \,\mathrm{d}\mathbb{P} \\
&=\; \big\|\mathbb{Z}_\star - \psi(X)\big\|^2_{L^2(\mathcal{H}_\mathcal{Y})} \;\leq\; \varepsilon^2, \quad \text{where} \quad \mathcal{A}_R \coloneqq \{\|\psi(X)\| \;\leq\; 2R\},
\end{aligned}$$

since $\mathcal{A}_R^c \subseteq \big\{\omega\in\Omega \mid \|\mathbb{Z}_\star(\omega)\| \leq R = 2R - R < \big|\|\psi(X(\omega))\| - \|\mathbb{Z}_\star(\omega)\|\big| \leq \|\psi(X(\omega)) - \mathbb{Z}_\star(\omega)\|\big\} \cup \mathcal{N}_\star$ for the $\mathbb{P}$-nullset $\mathcal{N}_\star \coloneqq \{\|\mathbb{Z}_\star\| > R\}$. This proves (ii), as desired.

As to (iii), note that clearly $\mathrm{ev}_X(\mathcal{B}_R) \subseteq \mathfrak{C}_R$, so let us fix any $\mathbb{Z} \in \mathfrak{C}_R$ for the converse inclusion. Then $\mathbb{Z}$ is $(\Sigma_X, \mathcal{B}(\mathcal{H}_\mathcal{Y}))$-measurable by definition of $\mathfrak{C}_R$, and hence, by Doob-Dynkin, $\mathbb{Z} = \tilde{\phi}_0(X)$ for some $(\mathcal{B}(\mathfrak{D}_X), \mathcal{B}(\mathcal{H}_\mathcal{Y}))$-measurable function $\tilde{\phi}_0 : \mathfrak{D}_X \to \mathcal{H}_\mathcal{Y}$; see e.g. [23, Lemma 1.14]. (Note that this $\tilde{\phi}_0$ is unique $\mathbb{P}_X$-a.e.: if $\mathbb{Z} = \tilde{\phi}_0(X) = \tilde{\phi}_1(X)$ then $\Delta \coloneqq (\tilde{\phi}_0 - \tilde{\phi}_1)\circ X = 0$ and thus $1 = \mathbb{P}(\Delta^{-1}\{0\}) = \mathbb{P}_X(\tilde{\phi}_0 = \tilde{\phi}_1)$.) Denoting $B_R \coloneqq \{\boldsymbol{t} \in \mathcal{H}_\mathcal{Y} \mid \|\boldsymbol{t}\| \leq R\} \in \mathcal{B}(\mathcal{H}_\mathcal{Y})$ and $\hat{\Omega} \coloneqq \mathbb{Z}^{-1}(B_R) \in \mathscr{F}$ and $\tilde{\mathbb{Z}} \coloneqq \mathbb{1}_{\hat{\Omega}} \cdot \mathbb{Z}$, note that ($\mathbb{Z} = \tilde{\mathbb{Z}}$ $\mathbb{P}$-a.s. by definition of $\mathfrak{C}$ and thus) $\mathbb{Z} = \tilde{\mathbb{Z}}$ in $L^2(\mathbb{P};\mathcal{H}_\mathcal{Y})$. Hence, and since $\tilde{\mathbb{Z}} = \phi(X)$ for the map $\phi \equiv \phi(x) \coloneqq \mathbb{1}_{\tilde{\phi}_0^{-1}(B_R)}(x) \cdot \tilde{\phi}_0(x)$ from $\mathcal{B}_R$, we have the identity $\mathbb{Z} = \phi(X)$ (in $L^2(\mathbb{P};\mathcal{H}_\mathcal{Y})$). This proves (iii), as claimed.

Statement (iv) follows via the conditional Jensen's inequality, as stated in [57], which gives that

$$\big\|\mathbb{E}\big[\mathbb{Y}^\Lambda \,\big|\, X\big]\big\|_{\mathcal{L}^\infty} \;\leq\; \operatorname{ess\,sup}_{\omega\in\Omega} \mathbb{E}\big[\|\mathbb{Y}^\Lambda\| \,\big|\, X\big](\omega) \;\leq\; R_\Lambda$$

(the last inequality holds by (48) and the monotonicity of the operator P_G from (139)); [57, $f = \|\cdot\|$] is applicable since $\mathbb{Y}^\Lambda$ is Bochner-integrable by [its inclusion in (54) and] [58, Proposition 1.16]. □

Given $R \geq 0$ and with notation (58), denote $\psi_{\alpha|R} \coloneqq \psi_\alpha \cdot \mathbb{1}_{\{\|\psi_\alpha\|\leq 2R\}}$ for $\alpha \in \mathfrak{L}^2_X$.

Proposition B.10. *In the setting of Theorem 4.8, and with* $\psi_{\alpha|R} \coloneqq \psi_\alpha \cdot \mathbb{1}_{\{\|\psi_\alpha\|\leq 2R\}}$ *for* $\alpha \in \mathfrak{L}^2_X$,

$$\mathbb{E}\big[\mathbb{Y}^\Lambda \,\big|\, X\big] \;=\; \operatorname*{arg\,min}_{\mathbb{Z}\,\in\,\mathfrak{C}_{R_\Lambda}} \Phi(\mathbb{Z}) \;=\; \lim_{k\to\infty} \psi_{\alpha_k|R_\Lambda}(X) \quad \textit{in} \;\; L^2_X(\mathcal{H}_\mathcal{Y}) \tag{168}$$

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

$$\inf_{\alpha\in\mathfrak{L}^2_X} \mathbb{E}\big[\|\mathbb{Y}^\Lambda - \psi_{\alpha|R_\Lambda}(X)\|^2\big]. \tag{169}$$

The convergence in (168) *holds* $\mathbb{P}$*-a.s. if* (α_k) *is such that* $\sum_{k=0}^\infty (\Phi(\psi_{\alpha_k|R_\Lambda}(X)) - \gamma_{R_\Lambda})^{1/2} < \infty$ *for* $\gamma_{R_\Lambda} \coloneqq \inf_{\alpha\in\mathfrak{L}^2_X} \Phi(\psi_{\alpha|R_\Lambda}(X))$.

Proof. The above are straightforward consequences of Lemma B.9 and the proof of Theorem 4.8.

Indeed: Adopting the notation of said proof, we see that

$$\mathbb{Y}_X^\Lambda \;=\; \operatorname*{arg\,min}_{\mathbb{Z}\,\in\,\mathfrak{C}_{R_\Lambda}} \mathbb{E}\Big[\big\|\mathbb{Y}^\Lambda - \mathbb{Z}\big\|^2\Big] \;=\; \operatorname*{arg\,min}_{\mathbb{Z}\,\in\,\mathfrak{C}_{2R_\Lambda}} \Phi(\mathbb{Z}) \tag{170}$$

from the characterisation (141) and the inclusions $\mathfrak{C}_{2R_\Lambda} \subset G$ and $\mathbb{Y}_X^\Lambda \in \mathfrak{C}_{R_\Lambda}$ (Lemma B.9 (iv)).

The convergence (168) then follows by combination of (170) and Lemma B.9 (ii), in complete analogy to how (65) was obtained by combination of (141) and (142): Noting that

$$\Phi(\mathbb{Y}_X^\Lambda) \overset{(170)}{=} \min_{\mathbb{Z}\in\mathfrak{C}_{R_\Lambda}} \Phi(\mathbb{Z}) \overset{\text{Lem. B.9 (ii)}}{=} \inf\{\Phi(\mathbb{Z}) \mid \mathbb{Z} \in \Psi_{2R_\Lambda}(X)\} \;=\; \inf_{\alpha\in\mathfrak{L}_X^2} \Phi(\psi_{\alpha|R_\Lambda}(X)) \;=\; \gamma_{R_\Lambda},$$

we see that, for any minimizing sequence (α_k) of $\{\Phi(\psi_{\alpha|R_\Lambda}(X)) \mid \alpha \in \mathfrak{L}_X^2\}$, the sequence $(\mathbb{Z}_k) := (\psi_{\alpha_k|R_\Lambda}(X)) \subset \mathfrak{C}_{2R_\Lambda}$ is minimizing for $\min\{\Phi(\mathbb{Z}) \mid \mathbb{Z} \in \mathfrak{C}_{2R_\Lambda}\}$; hence and with Lemma B.9 (i), both the convergence in (168) and the proposition's final almost-sure assertion follow exactly as in the proof of Theorem 4.8. □