

Orthogonal polynomials on path-space

Ilya Chevyrev¹, Emilio Ferrucci¹, Darrick Lee², Terry Lyons^{3,4}, Harald Oberhauser³, Nikolas Tapia⁵

submitted: February 23, 2026

<p>¹ SISSA (International School for Advanced Studies) via Bonomea 265 34136 Trieste Italy E-Mail: ichevyrev@gmail.com emilio.ferrucci@sissa.it</p>	<p>² School of Mathematics and Maxwell Institute University of Edinburgh Edinburgh EH9 3FD UK E-Mail: darrick.lee@ed.ac.uk</p>
---	---

<p>³ Mathematical Institute University of Oxford Woodstock Rd Oxford OX2 6GG UK E-Mail: terry.lyons@maths.ox.ac.uk oberhauser@maths.ox.ac.uk</p>	<p>⁴ Department of Mathematics Imperial College London 180 Queen's Gate London SW7 2AZ UK E-Mail: terry.lyons@maths.ox.ac.uk</p>	<p>⁵ Weierstrass Institute Anton-Wilhelm-Amo-Str. 39 10117 Berlin, Germany and Institut für Mathematik Humboldt-Universität zu Berlin Rudower Chaussee 25 12486 Berlin, Germany E-Mail: nikolasesteban.tapiamunoz@wias-berlin.de</p>
--	---	---

No. 3261

Berlin 2026



2020 *Mathematics Subject Classification.* 60L10, 42C05.

Key words and phrases. Expected signature, orthogonal polynomials.

We thank Cris Salvi for many helpful discussions on the topics of this paper. IC and EF gratefully acknowledge support from the ERC via the Starting Grant SQGT 101116964. During the first phase of this project, EF was employed at the University of Oxford and supported by the EPSRC programme grant [EP/S026347/1]. DL was supported by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA) during part of this work. TL was supported in part by UK Research and Innovation (UKRI) through the Engineering and Physical Sciences Research Council (EPSRC) via Programme Grants [Grant No. UKRI1010: High order mathematical and computational infrastructure for streamed data that enhance contemporary generative and large language models], [Grant No. EP/S026347/1: Unparameterised multi-model data, high order signatures and the mathematics of data science], and the UKRI AI for Science award [Grant No. UKRI2385: Creating Foundational Benchmarks for AI in Physical and Biological Complexity]. He was also supported by The Alan Turing Institute under the Defence and Security Programme (funded by the UK Government) and through the provision of research facilities; by the UK Government; and through CIMDA@Oxford, part of the AIR@InnoHK initiative funded by the Innovation and Technology Commission, HKSAR Government. HO was supported by [EP/Y028872/1] (An "Erlangen Programme" for AI) and by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA). NT acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – CRC/TRR 388 "Rough Analysis, Stochastic Dynamics and Related Fields" – Project ID 516748464. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) license to any author accepted manuscript version arising from this submission.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Anton-Wilhelm-Amo-Straße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: <http://www.wias-berlin.de/>

Orthogonal polynomials on path-space

Ilya Chevyrev, Emilio Ferrucci, Darrick Lee, Terry Lyons, Harald Oberhauser, Nikolas Tapia

ABSTRACT. We consider the orthogonalisation of the signature of a stochastic process as the analogue of orthogonal polynomials on path-space. Under an infinite radius of convergence assumption, we prove density of linear functions on the signature in L^p functions on grouplike elements, making it possible to represent a square-integrable function on (rough) paths as an L^2 -convergent series. By viewing the shuffle algebra as commutative polynomials on the free Lie algebra, we revisit much of the theory of classical orthogonal polynomials in several variables, such as the recurrence relation and Favard's theorem. Finally, we restrict our attention to the case of Brownian motion with and without drift, and prove that dimension-independent orthogonal signature exists with drift but not without. We end with numerical examples of how orthogonal signature polynomials of Brownian motion can be applied for the approximation of functions on paths sampled from the Wiener measure.

1. INTRODUCTION

Orthogonal polynomials in one or several variables have applications in numerical analysis [59], mathematical physics [51], probability theory and stochastic processes [55, 60], mathematical finance [1], and in many other areas of mathematics and the applied sciences. One simple example that exhibits their potential is the following. Take a function $f: [-1, 1] \rightarrow \mathbb{R}$. If f is analytic at 0, we may approximate it with its Taylor polynomials near 0. Outside its radius of convergence, however, the Taylor approximation will not converge, and even within, it may converge slowly. On the other hand, as long as $f \in L^2[-1, 1]$, its L^2 -projections onto the space spanned by polynomials of degree N , $\Pi_N f$, will converge to f in L^2 as $N \rightarrow \infty$ (and even uniformly if f has Hölder regularity greater than $\frac{1}{2}$ [53]). These projections can be found computing the L^2 inner product $(\ell_n, f) = \int_{-1}^1 f(x)\ell_n(x)dx$ with the Legendre polynomials, orthogonal w.r.t. the Lebesgue measure on $[-1, 1]$ and expanding

$$f = \sum_{n=0}^{\infty} \frac{(\ell_n, f)}{(\ell_n, \ell_n)} \ell_n. \quad (1.1)$$

The series truncated at N coincides with $\Pi_N f$. A comparison of the two types of approximation is given in Figure 1.

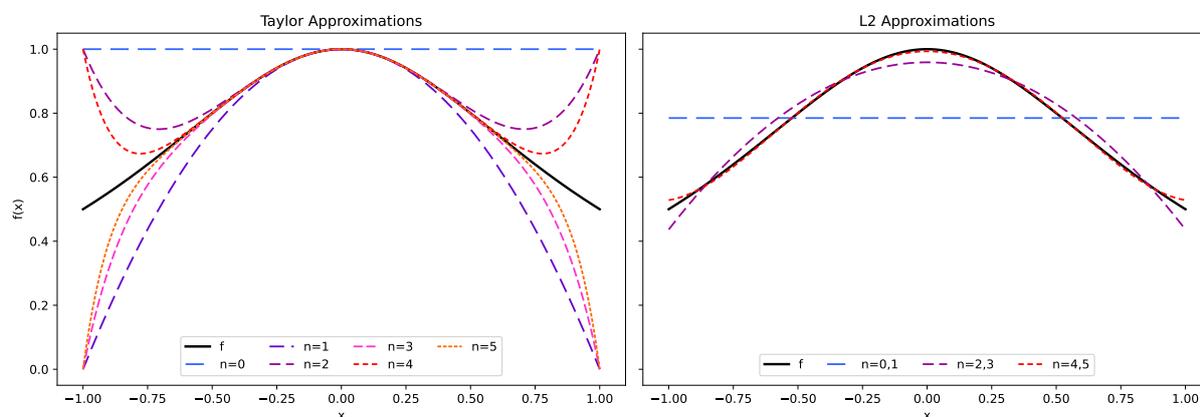


FIGURE 1. Taylor and L^2 approximations (computed with Legendre orthogonal polynomials) of $f(x) = \frac{1}{1+x^2}$. Since f has poles at $\pm i$, the Taylor approximations do not converge uniformly on $(-1, 1)$.

Since the work of Chen [15], iterated path integrals

$$S^n(X)_{0,T} := \int_{0 < t_1 < \dots < t_n < T} dX_{t_1} \otimes \dots \otimes dX_{t_n}, \quad X: [0, T] \rightarrow \mathbb{R}^d$$

have established themselves as the path-space analogue of polynomials. For rougher signals they are ill-defined from analysis alone, but in some cases can be defined probabilistically; specifying the first few subject to certain requirements forms the definition of a *rough path* [42], i.e., a specification of a theory of differential equations. The full stack of iterated integrals, its *signature* $S(X)_{0,T}$, has been shown to characterise the path up to *tree-like equivalence*, a generalised form of reparametrisation which allows “retracings”, see [11, 16, 33] for the proof of this statement in the case of X piecewise regular, bounded variation, and rough, respectively. Moreover, since coordinate functions on the signature form an algebra under the shuffle product (see (2.3) below), the Stone-Weierstrass theorem guarantees that, roughly speaking, for any

continuous function on a compact set of (rough) paths K and any $\varepsilon > 0$ there exists a truncation level N and a linear function λ_N of $S_N(X)_{0,T}$ (the signature truncated at degree N , see (2.2) below), such that

$$\sup_{X \in K} |F(X) - \langle \lambda_N, S_N(X)_{0,T} \rangle| < \varepsilon. \quad (1.2)$$

This result proves that functions on paths can be approximately linearised in terms of the signature, but has a couple of shortcomings. First of all, most sets of paths on which interesting functions are defined (e.g. the support of the law on path-space of the solution to an SDE [56], or indeed the support of the law of Brownian motion itself) are not compact. Secondly, writing $\lambda_N = \sum_{k=0}^N \lambda_N^k$ with λ_N^k acting on tensor degree k , it cannot be expected that a linear function $\lambda_{N'}$ ($N' > N$) that achieves a better approximation ($\varepsilon' < \varepsilon$) should have the same low-order tensor projections: in general $\lambda_N^k \neq \lambda_{N'}^k$ for $k \leq N$, i.e. the approximation is not stable under increasing the precision of the approximation. This latter phenomenon is not specific to approximations of the type (1.2) for signatures and even occurs for L^2 -approximations of functions in one variable with polynomials: if $f \in L^2[-1, 1]$ as above, writing $(\Pi_N f)(x) = \sum_{k=0}^N \lambda_N^k x^k$, the same instability occurs, simply because monomials are correlated w.r.t. the Lebesgue measure on $[-1, 1]$ (or indeed w.r.t. any measure that is not the Dirac delta δ_0).

Iterated integrals also naturally appear in numerical schemes for stochastic differential equations (SDEs) [36], see also [27, 40] for similar schemes which deal with path-dependency. These approximations are quite different to (1.2), since they generally rely on the interval $[0, T]$ being split into many sub-intervals, and on the expansion to be performed on each interval, for convergence to be guaranteed. More specifically, an important example of function on paths is $\varphi(Y_T)$, where φ is a smooth function and Y is the solution to the controlled differential equation with smooth vector fields $F_1, \dots, F_d \in C^\infty(\mathbb{R}^n, \mathbb{R}^n)$:

$$dY_t = F(Y_t) d\mathbf{X}_t, \quad Y_0 = y_0. \quad (1.3)$$

Applying the chain rule N times yields the expansion

$$\begin{aligned} \varphi(Y_T) &= \varphi(y_0) + \sum_{n=1}^N \langle F_{\alpha_1} \cdots F_{\alpha_n} \varphi(y_0), S(X)_{0,T}^{\alpha_1 \dots \alpha_n} \rangle + R_N(F, \varphi, X)_{0,T} \\ \text{with } R_N(F, \varphi, X)_{0,T} &= \int_{0 < t_1 < \dots < t_N < T} F_{\alpha_1} \cdots F_{\alpha_N} \varphi(Y_{t_1}) dX_{t_1}^{\alpha_1} \cdots dX_{t_N}^{\alpha_N}. \end{aligned} \quad (1.4)$$

Failures of analyticity such as Figure 1, which a fortiori occur in the path setting (since iterated integrals embed monomials) show that this error cannot generally be made small without controlling the size of the interval. This type of approximation scheme therefore has the benefit of having explicit identities for the coefficients in the expansion, but the drawbacks of requiring a more explicit form of the function on paths (e.g. the vector fields in the differential equation) and of requiring iteration for convergence.

Briefly returning to a polynomial L^2 -approximations for functions in one variable, observe that a necessary condition for a representation like (1.1) to hold is that polynomials be dense in L^2 . This is always true if the measure is compactly supported, but for general measures on the real line the question of density of polynomials in L^p is intimately related to that of moment-determinacy [9], i.e. whether the measure is unique given its moments $\mathbb{E}_\mu x^n$. On path-space, the analogous question asks whether the *expected signature* $\mathbb{E}S(X)_{0,T}$ of a stochastic process determines its law. The question was answered in the affirmative in [19] under an “infinite radius of convergence” assumption, which holds true for many stochastic processes lifted to rough paths considered in the literature [30].

The goal of this paper is to combine orthogonal polynomials with path signatures, forming what we consider the analogue of orthogonal polynomials on path-space. Our contributions are as follows. In Section 2 we briefly introduce the signature of a stochastic process and show how its canonical coordinates can be (block-)orthogonalised with the familiar Gram-Schmidt procedure, given the knowledge of the expected signature of the process. Next, we prove that, under the same infinite radius of convergence assumption as in [19], linear functions on the signature are dense in L^p functions on paths (modulo tree-like equivalence) for any $p \in [1, \infty)$. While this result holds under the same hypotheses as in [19], it does not follow directly from the latter. Combined with orthogonalisation, our density result yields an L^2 -series expansion for a square integrable function on paths F . This circumvents both the issues of poor convergence away from the point of expansion of (1.4) and the issues of stability under increase in precision of (1.2). We remark that L^p -density of signature functionals was also very recently proved in [14] using weighted spaces for a class of time-augmented processes. Our

density result is, however, stronger¹ and applies without any assumptions beyond an infinite radius of convergence. Our proof is moreover very different from [14] and follows more the style of the proof of moment determinacy [19].

In Section 3 we begin by observing the known fact that a change of basis transforms the shuffle algebra, which governs the product of coordinate functions on the signature, into the polynomial algebra over an infinite-dimensional but graded space (the free Lie algebra). This enables us to revisit much of the theory of classical orthogonal polynomials in several variables [25], with the added difficulty of infinite dimensionality and anisotropic grading on the base variables. We prove versions of the celebrated three-term recurrence relation, which is no longer three-term but nevertheless yields a way of computing the orthogonal signature polynomials alternatively to Gram-Schmidt. We also prove a version of Favard's theorem, which characterises which sequences of polynomials arise as orthogonal polynomials with respect to an inner product. We then study when such inner products arise from probability measures on both the free Lie algebra and path space.

In Section 4 we specify our study to what is arguably the most important example of stochastic process, Brownian motion, seeking more explicit descriptions of the orthogonal signature polynomials. We ask the question of *naturality*, i.e. roughly speaking whether the orthogonal polynomials associated to a V -valued brownian motion can be expressed without reference to a basis or even to the dimension of V . While this is true of ordinary Hermite polynomials in several variables, we prove, using symbolic linear algebra code that it is not true for the signature of Brownian motion, unless time is included as a coordinate. This concern is not merely motivated by aesthetics, rather, it speeds up the computation of the orthogonal signature compared to naively performing Gram-Schmidt orthogonalisation. We proceed with the explicit description and computation of orthogonal signature polynomials of time-augmented Brownian motion which are natural (basis- and dimension- free). We also give an independent, short, Wiener chaos-based proof of density of linear functions on the signature in L^2 in the time-augmented case. We end with a description of our implementation [17] of these orthogonal polynomials on Wiener space and show how it can be used in approximation problems.

We believe that, in future work, it would be interesting to consider the orthogonal signature associated to the time-augmented stochastic processes (many of which with jumps) associated to the Askey classification [3], see [55].

2. L^2 ORTHOGONAL SIGNATURE EXPANSIONS

2.1. Signature of a stochastic process. Let X be a continuous stochastic process in a time variable on $[0, T]$ taking values in the finite-dimensional vector space V , defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let X be of bounded p -variation and let \mathbf{X} be a p -geometric rough path lift of X , which we assume to be \mathcal{F} -measurable (this is the case for all the main examples of stochastic rough paths). For a finite-dimensional vector space U we denote

$$T_N(U) := \bigoplus_{m=0}^N U^{\otimes m}, \quad T(U) := \bigoplus_{m=0}^{\infty} U^{\otimes m}, \quad \mathcal{T}(U) := \prod_{m=0}^{\infty} U^{\otimes m}. \quad (2.1)$$

We recall that, given the rough path lift \mathbf{X} , its *signature* is canonically defined [42], and represents the iterated integrals

$$\begin{aligned} S^n(\mathbf{X})_{0,T} &= \int_{0 < t_1 < \dots < t_n < T} dX_{t_1} \otimes \dots \otimes dX_{t_n} \in V^{\otimes n} \\ S_N(\mathbf{X})_{0,T} &= (S^0(\mathbf{X})_{0,T}, \dots, S^N(\mathbf{X})_{0,T}) \in T_N(V) \\ S(\mathbf{X})_{0,T} &= (S^0(\mathbf{X})_{0,T}, \dots, S^N(\mathbf{X})_{0,T}, \dots) \in \mathcal{T}(V) \end{aligned} \quad (2.2)$$

defined according to the integration theory specified by \mathbf{X} . We write $\langle \cdot, \cdot \rangle$ for the canonical dual pairing between $T(V^*)$ and $\mathcal{T}(V)$, e.g. for a word $w = \alpha_1 \dots \alpha_n$ representing an elementary tensor in $(V^*)^{\otimes n}$ we have

$$\langle w, S(\mathbf{X})_{0,T} \rangle := \int_{0 < t_1 < \dots < t_n < T} dX_{t_1}^{\alpha_1} \otimes \dots \otimes dX_{t_n}^{\alpha_n}.$$

We write $|w| = n$ for the tensor degree of w . We recall that signature coordinates satisfy the *shuffle identity*, i.e. for $u, v \in \mathcal{T}(V^*)$

$$\langle u, S(\mathbf{X}) \rangle \langle v, S(\mathbf{X}) \rangle = \langle u \sqcup v, S(\mathbf{X}) \rangle. \quad (2.3)$$

where \sqcup is the shuffle product of the two words. For this reason we sometimes will call $\text{Sh}(V^*) = \mathcal{T}(V^*)$ when viewed as an algebra under the shuffle product and correspondingly write $\text{Sh}_N(V^*) = T_N(V^*)$. Elements of $\mathcal{T}(V)$ satisfying (2.3) form

¹More specifically, [14] requires time-augmentation and integrability of the weight $w(\mathbf{X}) = \exp(\beta \|\mathbf{X}\|_\alpha^N)$ for some $\beta > 0$, where $N = \lfloor 1/\alpha \rfloor$ and $\|\mathbf{X}\|_\alpha$ is the α -Hölder homogeneous rough path norm. Integrability of w (for any $\beta > 0$) implies an infinite radius of convergence of $\mathbb{E}S(\mathbf{X})_{0,T}$, so our Theorem 2.4 implies [14, Thm. 3.4], but not conversely since we do not require time augmentation and do not assume integrability of w . For example, Theorem 2.4 applies to fractional Brownian motion with $1/4 < H \leq 1/3$ while [14, Thm. 3.4] does not because w is not integrability as in this case $N = 3$. Moreover, the restriction to time augmented processes seems important in [14] since the proof relies on the uniform density result in weighted spaces from [23], which appears to use time augmentation in a crucial way.

a (formal) group $G(V) \subset T(V)$, whose Lie algebra is the space of free Lie series [50]. Recall that $A = (A^0, A^1, \dots) \in T(V)$ has an infinite radius of convergence if

$$\|A\|_\lambda := \sum_{n=0}^{\infty} \|A^n\|_{V^{\otimes n}} \lambda^n < \infty \quad \text{for all } \lambda > 0, \quad (2.4)$$

where we equip V with a norm and $V^{\otimes n}$ with an arbitrary cross norm, e.g. the projective norm (by finite dimensionality of V , condition (2.4) does not depend on the choice of norms). Denote $\tilde{T}(V)$ the space of these elements and

$$\tilde{G}(V) := \tilde{T}(V) \cap G(V).$$

We equip $\tilde{T}(V)$ with the (locally convex) topology given by the norms $\|\cdot\|_\lambda$, and $\tilde{G}(V)$ inherits this topology, making it a topological group, and a Polish (completely metrisable and separable) space [19].

It is a fact that $S(X)_{0,T}$ is an element of $\tilde{G}(V)$ for any rough path X , see e.g. [43]. Therefore, we may view the stochastic process X as defining a probability measure on $\tilde{G}(V)$. From now on, unless there is an ambiguity as to the rough path lift X , we will drop the bold font and simply write $S(X)_{0,T}$ and similar; we will always use X to denote the trace of X , i.e. its projection onto V .

The signature encodes almost all the information in the (rough) path, namely up to tree-like equivalence [11, 16, 33], a generalised form of reparametrisation that includes “retracings”. A common way of ensuring the path is tree-reduced, and thus that the full path can be recovered from the signature, is to include time as a coordinate, see e.g. [29]. We let \mathcal{G} denote the sigma-algebra generated by $S(X)_{0,T}$.

Remark 2.1 ($\mathcal{G} = \mathcal{F}$). If time is included as zero-th coordinate of X , it follows from the fact that X is tree-reduced (i.e. $t \mapsto X_{0,t}$ is injective) and measurability of the map that reconstructs the tree-reduced representative from the signature [32], that $\mathcal{G} = \mathcal{F}$. In many cases of interest, such as Brownian motion [37] and the law of certain multidimensional diffusions [31], $\mathcal{G} = \mathcal{F}$ even without including time as a coordinate of the process.

2.2. Orthogonalisation of signature features. We assume that $\langle w, S(X)_{0,T} \rangle \in L^2(\mathcal{G})$ for all $w \in \text{Sh}(V^*)$. Define the positive semi-definite symmetric bilinear form

$$(u, v) := \mathbb{E} \langle u, S(X)_{0,T} \rangle \langle v, S(X)_{0,T} \rangle = \langle u \sqcup v, \mathbb{E} S(X)_{0,T} \rangle. \quad (2.5)$$

This bilinear form is, in general, only positive semi-definite because the natural map $\phi: \text{Sh}_N(V^*) \rightarrow L^2(\mathcal{G})$ is not in general an inclusion. Let N be its nullspace and W a choice of a direct complement to it, so that $\text{Sh}(V^*) = N \oplus W$ and ϕ embeds W into $L^2(\mathcal{G})$. We may always choose W to be of the form $W = \bigoplus_{n=0}^{\infty} W^n$ with $W^n \subset (V^*)^{\otimes n}$: indeed, let \mathcal{W}^n be the pre-image through the quotient $\pi: \text{Sh}(V^*) \rightarrow \text{Sh}(V^*)/N$ of some basis of $\pi((V^*)^{\otimes n})$ and take $W^n = \text{span}(\mathcal{W}^n)$. As usual call $\mathcal{W}_n = \bigsqcup_{k=0}^n \mathcal{W}^k$ the basis of $W_n = \bigoplus_{k=0}^n W^k$. Let

$$\begin{aligned} \Pi_N: L^2(\mathcal{G}) &\rightarrow \text{span}\{\langle \ell, S(X)_{0,T} \rangle \mid \ell \in \text{Sh}_N(V^*)\} \\ \Pi^n: L^2(\mathcal{G}) &\rightarrow \text{span}\{\langle \ell, S(X)_{0,T} \rangle \mid \ell \in (V^*)^{\otimes n}\} \\ \mathbb{1} - \Pi_N =: \Pi_N^\perp: L^2(\mathcal{G}) &\rightarrow \text{span}\{\langle \ell, S(X)_{0,T} \rangle \mid \ell \in \text{Sh}_N(V^*)\}^\perp \end{aligned} \quad (2.6)$$

denote orthogonal projections, where \perp denotes the orthogonal complement in $L^2(\mathcal{G})$. We consider the *block-orthogonalisation map* p

$$p: W \xrightarrow{\cong} W, \quad p|_{W^n} = \Pi_{n-1}^\perp|_{W^n}, \quad w \mapsto p_w. \quad (2.7)$$

Notice that $p|_{W^n}$ takes values in $W_{n-1} \oplus W^n = W_n$ and therefore so does $p|_{W_n}$, i.e. p is triangular, and moreover monic, namely $\Pi^n \circ p|_{W^n} = \mathbb{1}$. It can be computed by solving, for $w \in \mathcal{W}^n$ (and extended linearly to W^n)

$$p_w = w - \sum_{u \in \mathcal{W}_{n-1}} \lambda_u u, \quad \left(w - \sum_{u \in \mathcal{W}_{n-1}} \lambda_u u, v \right) = 0 \quad \text{for } v \in \mathcal{W}_{n-1}. \quad (2.8)$$

Let \mathcal{V} be a basis of V^* , then the words $\alpha_1 \dots \alpha_n$ in \mathcal{V}^n define a basis of $\text{Sh}(V^*)$, and write $\mathcal{V}^\bullet := \bigsqcup_{n \in \mathbb{N}} (\mathcal{V}^*)^{\otimes n}$. If the bilinear form (\cdot, \cdot) is positive-definite, then $W = \text{Sh}(V^*)$; in many cases it is possible to realise W^n as the linear span of a subset of \mathcal{V}^\bullet (see Section 4.2 below for an explicit example of this). Assume this is the case, and that \mathcal{W}^n is a subset of words in \mathcal{V}^n . We then call the elements $\{p_w\}_{w \in \mathcal{W}}$ the *block-orthogonal shuffle polynomials* since $(p_v, p_w) = 0$ for $v, w \in \mathcal{W}$ with $|v| \neq |w|$.

Assume furthermore the basis \mathcal{W} is totally ordered in a way that respects the grading ($v < w$ if $|v| < |w|$). We may then define a fully orthogonal basis by the Gram-Schmidt orthogonalisation procedure:

$$\tilde{p}_w := w - \sum_{\substack{v \in \mathcal{W}^\bullet \\ v < w \\ (\tilde{p}_v, \tilde{p}_v) > 0}} \frac{(w, \tilde{p}_v)}{(\tilde{p}_v, \tilde{p}_v)} \tilde{p}_v \quad (2.9)$$

where $v < w$ if $|v| < |w|$ or if $|v| = |w|$ and $v < w$ lexicographically. Then $\{\tilde{\rho}_w\}_{w \in \mathcal{W}^n}$ is an orthogonal basis of $\text{span}\{\rho_w \mid w \in \mathcal{W}^n\}$.

In the next subsection we will identify a large class of stochastic processes for which linear functions on $S(X)_{0,T}$ are dense in $L^p(\mathcal{G})$. We view the next two results as corollaries of that result.

Corollary 2.2 (Learning $F(X)$ by linear regression). *Let $Y = F(X) \in L^2(\mathcal{G})$ and $\{(X^i, Y^i = F(X^i) + \varepsilon^i)\}_{i=1}^M$ be an i.i.d. sample of input-output pairs, where ε^i are i.i.d. errors independent of X . Let $\Phi \in \mathbb{R}^{M \times D}$ be the data matrix with*

$$\Phi_{iv} = \langle \rho_v, S_N(X)_{0,T} \rangle, \quad i = 1, \dots, M; v \in \mathcal{V}^n$$

with $n \leq N$ and $D = D(d, N) = \frac{d^{N+1}-1}{d-1}$ the dimension of $T_N(\mathbb{R}^d)$. Then the estimator for ordinary least squares (OLS)

$$\hat{\beta} = (\Phi^T \Phi)^{-1} \Phi^T Y = \Phi^T (\Phi \Phi^T)^{-1} Y \tag{2.10}$$

is such that $\langle \hat{\beta}, S_N(X)_{0,T} \rangle$ converges in L^2 to $\Pi_N Y$ as $M \rightarrow \infty$. Its entries converge (in blocks) to $\Pi^N Y$, and are thus asymptotically stable under increasing degree N . Therefore, under the hypotheses of Theorem 2.4 below, Y can be estimated arbitrarily well by OLS linear regression on the truncated signature.

We refer to the standard literature on statistical learning (e.g. [5, Ch. 3]) for more general statements that imply its proof. In most real-world cases in which data is limited it is desirable or even necessary to use ridge or lasso regression; consistency results could be formulated for these too, by letting the regularisation parameter vanish in the large data limit. Consistency of the OLS estimator and the statement about universality of OLS estimation of course holds true without any orthogonalisation, since the projection Π_N does not depend on it. However, as explained in Section 1, orthogonalisation is necessary to obtain a series representation. The following is classical.

Corollary 2.3 (Series expansion of an L^2 function on paths). *Under the hypotheses of Theorem 2.4 below, for $Y \in L^2(\mathcal{G})$*

$$Y = \sum_{n=0}^{\infty} \Pi^n Y = \sum_{v: \langle \tilde{\rho}_v, \tilde{\rho}_v \rangle > 0} \langle \ell_v^Y, S(X)_{0,T} \rangle, \quad \ell_v^Y := \frac{\mathbb{E}[Y \langle \tilde{\rho}_v, S(X)_{0,T} \rangle]}{\langle \tilde{\rho}_v, \tilde{\rho}_v \rangle} \tilde{\rho}_v$$

with Π^n defined in (2.6), $\tilde{\rho}_v$ in (2.9), and convergence of the series in L^2 .

Proof. $\Pi_N = \sum_{n=0}^N \Pi^n$ and by density $\|Y - \Pi_N Y\|_{L^2} \rightarrow 0$. The second equality follows by further decomposing Π^n into the projections onto the orthogonal directions spanned by the vectors $\tilde{\rho}_i$. \square

2.3. Density of linear functions on the signature in L^p . The goal of this section is to prove the following result. We equip \mathbb{R}^d with the ℓ^1 norm $\|(x_1, \dots, x_d)\| = \sum_{i=1}^d |x_i|$ and equip all tensor products, including $(\mathbb{R}^d)^{\otimes k}$, with the projective tensor norm.

Theorem 2.4. *Consider a probability measure μ on $\tilde{G}(\mathbb{R}^d)$ with infinite radius of convergence, i.e.*

$$\sum_{k=0}^{\infty} \lambda^k \|\mathbb{E}_{S \sim \mu} S^k\|_{(\mathbb{R}^d)^{\otimes k}} < \infty \quad \text{for all } \lambda > 0. \tag{2.11}$$

Then the space of linear coordinate functions on $\tilde{G}(\mathbb{R}^d)$ are dense in $L^p(\mu)$ for all $p \in [1, \infty)$.

The exact choice of norms on $(\mathbb{R}^d)^{\otimes k}$ for $k \geq 1$ is not important in Theorem 2.4 because if μ verifies (2.11) for one choice of norms, then it verifies (2.11) for all choices of cross norms.

As we discuss in Section 1, a related L^p -density result, which is more restrictive as it requires time to be included as a coordinate and a weight function to be integrable, has very recently been obtained in [14]. See also Proposition 4.10 below for a short Wiener chaos-based proof that applies in the case $p = 2$ and X is a time-augmented Brownian motion. Theorem 2.4 applies to the stochastic processes to which the moment-determinacy results of [19] applies to, such as Gaussian and Markovian rough paths. Note that, while the hypothesis of infinite radius of convergence is the same as in [19], L^p density does not follow from moment determinacy alone: this is even true in finite dimension [10] (only in the scalar case and $p \leq 2$ does this implication hold [25, p.69]). We also remark that a similar approximation result for the modified robust signatures [20] was obtained in [7].

For the proof, we require a few lemmas. In the following, let (Y, ρ) be a complete (not necessarily separable) metric space with a Borel probability measure μ .

Lemma 2.5. *The space of continuous bounded functions is dense in $L^p(\mu)$ for any $p \in [1, \infty)$.*

Proof. The span of indicator functions on open sets is dense in $L^p(\mu)$. But for $A \subset Y$ open, the continuous bounded function $f_n(y) = \{n \inf_{x \in Y \setminus A} \rho(x, y)\} \wedge 1$ converges to $\mathbf{1}_A$ pointwise, and thus in L^p by dominated convergence. \square

We say that a set of functions $\mathcal{A} \subset \mathbb{R}^Y$ *separates points* if for all distinct $x, y \in Y$ there exists $h \in \mathcal{A}$ such that $h(x) \neq h(y)$.

Lemma 2.6. *Suppose \mathcal{A} is an algebra of continuous bounded functions on Y that separates points and contains the constant functions. Suppose also that μ is Radon (which is automatic if Y is separable). Then \mathcal{A} is dense in $L^p(\mu)$ for any $p \in [1, \infty)$.*

Proof. There are several proofs, see e.g. [24] for an even stronger statement that drops the continuity assumption. We give here an elementary proof that uses only the classical Stone–Weierstrass theorem. We let $C > 0$ denote a sufficiently large universal constant.

Let $\overline{\mathcal{A}}$ be the closure of \mathcal{A} in $L^p(\mu)$. By Lemma 2.5, it suffices to show that if $f: Y \rightarrow \mathbb{R}$ is continuous with $\sup_{y \in Y} |f(y)| \leq 1$, then for every $\varepsilon > 0$, there exists $h \in \overline{\mathcal{A}}$ such that $\|f - h\|_{L^p(\mu)} < \varepsilon$. To this end, consider $\varepsilon \in (0, 1)$. Let $K \subset Y$ be a compact set such that $\mu(Y \setminus K) < \varepsilon$; such a set K exists because μ is Radon. By Stone–Weierstrass, there exists $g \in \mathcal{A}$ such that $\sup_{y \in K} |g(y) - f(y)| < \varepsilon$ (of course g depends on K and f). At this stage, we *don't* have $\|g - f\|_{L^p(\mu)}$ small because g might be large outside K .

Now consider the functions $h_\delta(y) = e^{-\delta g(y)^2} g(y)$ for small $\delta > 0$. Define

$$H(\delta) = \sup_{x \in \mathbb{R}} |e^{-\delta x^2} x|.$$

Then $H(\delta) \leq C\delta^{-1/2}$ and thus

$$\sup_{y \in Y} |h_\delta(y)| \leq H(\delta) \leq C\delta^{-1/2}.$$

Moreover $\sup_{y \in K} |g(y) - h_\delta(y)| \leq C\delta$ since $\sup_{y \in K} |g(y)| \leq 2$ and $e^{-\delta z^2} = 1 + O(\delta z^2)$ for $|z| \leq 2$. In particular, $\sup_{y \in K} |f(y) - h_\delta(y)| \leq \varepsilon + C\delta$ and thus

$$\int_K |f - h_\delta|^p d\mu \leq (\varepsilon + C\delta)^p.$$

On the other hand,

$$\int_{Y \setminus K} \{f^p + h_\delta^p\} d\mu \leq \varepsilon + \varepsilon H(\delta)^p \leq \varepsilon + C^p \varepsilon \delta^{-p/2}$$

Taking $\delta = \varepsilon^{1/p}$, we obtain

$$\|f - h_\delta\|_{L^p(\mu)}^p \leq (\varepsilon + C\varepsilon^{1/p})^p + \varepsilon + C^p \varepsilon^{1/2}.$$

It remains to show that $h_\delta \in \overline{\mathcal{A}}$. We write as a power series $h_\delta = \sum_{n=0}^{\infty} \frac{(-\delta)^n g^{2n+1}}{n!}$ and remark that, since g is bounded on Y , the series converges absolutely in L^p . Since $g^n \in \mathcal{A}$ for all $n \geq 1$, we obtain $h_\delta \in \overline{\mathcal{A}}$. \square

Proof of Theorem 2.4. Let A denote the space of matrix coefficients of unitary representations of $\widetilde{G}(\mathbb{R}^d)$ as in [19], i.e. A contains those functions $f: \widetilde{G}(\mathbb{R}^d) \rightarrow \mathbb{C}$ such that $f(S) = \langle M(S)u, v \rangle_{\mathbb{C}^N}$ for some $u, v \in \mathbb{C}^N$ and a linear map $M: \mathbb{R}^d \rightarrow \mathfrak{u}(N)$ for some $N \geq 1$, where $\mathfrak{u}(N)$ is the space of $N \times N$ skew-Hermitian matrices and where $M(S) = \sum_{k=0}^{\infty} M^{\otimes k} S^k$, which is a unitary $N \times N$ matrix. Then A is a \mathbb{C} -algebra of bounded continuous functions on $\widetilde{G}(\mathbb{R}^d)$ which, by [19, Theorem 4.8] (see also [39] for a different proof), separates points. Moreover A contains the constant functions and is closed under complex conjugation, so by Lemma 2.6, functions of the form $\Re f$ for $f \in A$ are dense in $L^p(\mu)$. It remains to show that every $f \in A$ can be approximated in $L^p(\mu)$ by (\mathbb{C} -valued) linear functions on $\widetilde{G}(\mathbb{R}^d)$.

Consider a map $M: \mathbb{R}^d \rightarrow \mathfrak{u}(N)$ and let $S = (1, S^1, S^2, \dots)$ be a $\widetilde{G}(\mathbb{R}^d)$ -valued random variable with law μ . Since $\sum_{k=0}^n M^{\otimes k} S^k$ is a linear function on $\widetilde{G}(\mathbb{R}^d)$ and $M(S) = \sum_{k=0}^{\infty} M^{\otimes k} S^k$, it suffices to show that

$$\lim_{n \rightarrow \infty} \sum_{k=n}^{\infty} \|M\|^k (\mathbb{E} \|S^k\|_{(\mathbb{R}^d)^{\otimes k}}^p)^{1/p} = 0. \quad (2.12)$$

For normed spaces E, F , we equip $E \oplus F$ with the ℓ^1 norm $\|(x, y)\| = \|x\|_E + \|y\|_F$. Consider now $n \geq 2$ and the n -th iterated coproduct $\Delta_n: T(\mathbb{R}^d) \rightarrow T(\mathbb{R}^d)^{\otimes n}$, which is the unique algebra morphism given by $\Delta_n(v) = v \otimes 1 \otimes \dots \otimes 1 + \dots + 1 \otimes \dots \otimes 1 \otimes v$ for $v \in \mathbb{R}^d$, where the number of terms is n (see [50, Sec. 1.4]). Note that $\Delta_2 = \Delta$ is the usual coproduct on $T(\mathbb{R}^d)$ which is dual to the shuffle product. In particular, for $v \in \mathbb{R}^d$,

$$\|\Delta_n v\|_{(\mathbb{R}^d)^{\otimes n}} = n \|v\|_{\mathbb{R}^d} \quad (2.13)$$

and thus, for $x \in (\mathbb{R}^d)^{\otimes k}$,

$$\|\Delta_n x\| \leq n^k \|x\|_{(\mathbb{R}^d)^{\otimes k}}. \quad (2.14)$$

Here, $\Delta_n x$ is an element in $\bigoplus_{k_1 + \dots + k_n = k} \bigotimes_{i=1}^n (\mathbb{R}^d)^{\otimes k_i}$, which we recall is equipped with the ℓ^1 norm of the corresponding projective tensor norms. The bound (2.14) follows from (2.13) and the fact that, for any normed algebra B and linear map $Q: \mathbb{R}^d \rightarrow B$ with operator norm $\|Q\|$, $Q^{\otimes k}: (\mathbb{R}^d)^{\otimes k} \rightarrow B$ has operator norm bounded above by $\|Q\|^k$ due to the choice

of projective norms on $(\mathbb{R}^d)^{\otimes k}$. More precisely, we apply this with $Q = \Delta_n$, for which $\|Q\| = n$ by (2.13), and $B = T(\mathbb{R}^d)^{\otimes n}$ equipped with the projective norm, where we equip $T(\mathbb{R}^d)$ with the ℓ^1 norm $\|(x^0, x^1, \dots)\| = \sum_{k \geq 0} \|x^k\|_{(\mathbb{R}^d)^{\otimes k}}$.

Therefore, for $p \geq 1$ an integer, using the duality between shuffle and Δ as in [50, Sec. 1.5] and the fact that S takes values in $G(\mathbb{R}^d)$,

$$\begin{aligned} \mathbb{E}\|S^k\|_{(\mathbb{R}^d)^{\otimes k}}^{2p} &= \mathbb{E}\left(\sum_{|w|=k} |\langle w, S^k \rangle|\right)^{2p} \leq d^{k(2p-1)} \sum_{|w|=k} \mathbb{E}\langle w, S^k \rangle^{2p} = d^{k(2p-1)} \sum_{|w|=k} \mathbb{E}\langle w^{\sqcup 2p}, S^{2pk} \rangle \\ &= d^{k(2p-1)} \sum_{|w|=k} \mathbb{E}\langle w^{\otimes 2p}, \Delta_{2p} S^{2pk} \rangle \leq d^{k(2p-1)} \|\Delta_{2p} \mathbb{E} S^{2pk}\| \\ &\leq d^{k(2p-1)} (2p)^{2pk} \|\mathbb{E} S^{2pk}\|_{(\mathbb{R}^d)^{\otimes 2pk}}, \end{aligned}$$

where w in the sums ranges over all words of length k in the canonical basis of \mathbb{R}^d (this generalises the calculation above [19, Proposition 3.4] which was for $p = 1$). Since $\mathbb{E}S$ has an infinite radius of convergence, for any $\varepsilon > 0$, one has $\|\mathbb{E}S^k\| \leq \varepsilon^k$ for all $k > k(\varepsilon) \geq 1$ sufficiently large. Therefore, for all $k > k(\varepsilon)$,

$$(\mathbb{E}\|S^k\|^{2p})^{1/(2p)} \leq d^{\frac{k(2p-1)}{2p}} (2p)^k \varepsilon^k.$$

Taking ε sufficiently small (depending on $\|M\|$, p , and d), we obtain (2.12) as required. \square

3. GENERAL PROPERTIES OF ORTHOGONAL SHUFFLE POLYNOMIALS

In this section, we discuss structural properties of orthogonal shuffle polynomials in $\text{Sh}(V^*)$, generalizing several classical results in the classical commutative setting. We begin by reformulating shuffle polynomials as ordinary commutative polynomials with graded generators.

3.1. Shuffle polynomials as graded commutative polynomials. For a vector space V , let $\text{Sym}(V)$ denote the symmetric algebra of V . Recall that $L(V)$ denotes the free Lie algebra of V and, from [50],

$$\text{Sh}(V^*) = \text{Sym}(L(V)^{\text{gr}}) \tag{3.1}$$

as commutative algebras, where $(\cdot)^{\text{gr}}$ denotes the graded dual and identity is intended as being between functors. Fixing a Hall basis of $L(V)$ allows us to express elements of $\text{Sh}(V^*)$ as polynomials over an infinite set of indeterminates, specified by the dual Hall basis [50, §5.2]. As many of the results in this section do not rely on the Lie algebra structure, we will mainly work with a general graded vector space $W = \bigoplus_{m=0}^{\infty} W^m$ of finite type, where the degree m subspace W^m is finite dimensional. Our general results for $\text{Sym}(W^{\text{gr}})$ apply to the signature setting by taking $W = L(V)$.

We fix a basis of W^m , denoted $\{w_{m,i}\}_{i=1}^{N_m}$, where N_m is the dimension of W^m . We abuse notation, and also denote the dual basis in $(W^m)^*$ by $w_{m,i}$. In general, we will denote a monomial in $\text{Sym}(W^{\text{gr}})$ by \mathbf{w}^α , where $\alpha = (\alpha_{m,i})_{m \in \mathbb{N}, i \in [N_m]}$, where each $\alpha_{m,i} \in \mathbb{N}$, and only finitely many are nonzero, and set

$$\mathbf{w}^\alpha = \prod_{m=1}^{\infty} \prod_{i=1}^{N_m} w_{m,i}^{\alpha_{m,i}}. \tag{3.2}$$

Let $Q \in \text{Sym}(W^{\text{gr}})$ be a polynomial $Q = \sum_{\alpha} q_{\alpha} \mathbf{w}^{\alpha}$, where finitely many of the q_{α} are nonzero. We have two notions of degree called the *tensor (total) degree* and the *shuffle degree*, respectively defined on a shuffle monomial \mathbf{w}^α by

$$\text{tdeg}(\mathbf{w}^\alpha) = |\alpha| = \sum_{m=1}^{\infty} \sum_{i=1}^{N_m} m \alpha_{m,i} \quad \text{and} \quad \text{sdeg}(\mathbf{w}^\alpha) = \sum_{m=1}^{\infty} \sum_{i=1}^{N_m} \alpha_{m,i}.$$

We denote the column vector of shuffle monomials of total degree n by \mathbf{w}_n and denote its dimension by r_n .

Definition 3.1. A linear functional $\mathcal{L} : \text{Sym}(W^{\text{gr}}) \rightarrow \mathbb{R}$ is *quasi-definite* if there exists a basis of $\text{Sym}(W^{\text{gr}})$ such that for any two basis elements P, Q , we have

$$\mathcal{L}(PQ) = 0 \text{ if } P \neq Q \quad \text{and} \quad \mathcal{L}(P^2) \neq 0.$$

If in addition, \mathcal{L} satisfies $\mathcal{L}(P^2) > 0$, then we say that \mathcal{L} is positive-definite.

Given a quasi-definite functional \mathcal{L} , define the symmetric bilinear form (or inner product, if \mathcal{L} is positive-definite)

$$(P, Q) := \mathcal{L}(PQ),$$

and say that $\{\mathbf{p}_n\}_{n \geq 0}$, where $\mathbf{p}_n \in \text{Sym}(W^{\text{gr}})^{r_n}$ is *block orthogonal* if $(\mathbf{w}_m, \mathbf{p}_n^T) = 0$ for all $m < n$ and $(\mathbf{w}_n, \mathbf{p}_n^T)$ is invertible. In this section, orthogonal polynomials refer to block orthogonal polynomials. We say that $\{\tilde{\mathbf{p}}_n\}_{n \geq 0}$ is *orthonormal* if in

addition, $(\tilde{\mathbf{p}}_n, \tilde{\mathbf{p}}_n^\top)$ is the identity. We note that there are two main differences from the classical orthogonal polynomial setting:

- 1 there are an infinite number of generators, ie. W is infinite-dimensional; and
- 2 the generators are *graded*, and thus orthogonalize with respect to the *total* degree.

In the following subsections, we generalize classical results about orthogonal polynomials [25] to our graded, infinite-dimensional setting. We show that orthogonal polynomials on $\text{Sym}(W^{\text{gr}})$ satisfy a recurrence relation along with rank conditions on the defining matrices. In Section 3.4, we prove the converse in a generalization of Favard's theorem, and furthermore discuss when inner products on $\text{Sym}(W^{\text{gr}})$ are induced by probability measures in Section 3.5. We return to the specific case of $W = L(V)$ in Section 3.6 and discuss measures on $G(V)$.

3.2. Recurrence relation. We begin by assuming that \mathcal{L} is a quasi-definite linear functional on $\text{Sym}(W^{\text{gr}})$ and that $\{\mathbf{p}_n\}_{n \geq 0}$ denotes a system of block orthogonal polynomials. Let

$$H_n = (\mathbf{p}_n, \mathbf{p}_n^\top) \in \text{Mat}(r_n, r_n)$$

which is symmetric and invertible by definition of block orthogonality. We begin by generalizing the three-term relation of [25, Theorem 3.3.1] to the graded setting.

Proposition 3.2. *For $n \in \mathbb{N}_0$, $m \in [n]$, $i \in [N_m]$ and $-m \leq k \leq m$, there exist unique matrices $M_{n,m,i}^k \in \text{Mat}(r_{n-m}, r_{n-m+k})$ such that*

$$w_{m,i} \mathbf{p}_{n-m} = \sum_{k=-m}^m M_{n,m,i}^k \mathbf{p}_{n-m+k}. \quad (3.3)$$

These matrices satisfy

$$M_{n,m,i}^k H_{n-m+k} = H_{n-m} (M_{n+k,m,i}^{-k})^\top. \quad (3.4)$$

Proof. The components of $w_{m,i} \mathbf{p}_{n-m}$ are polynomials of degree n , so they can be written as a linear combination of orthogonal polynomials,

$$w_{m,i} \mathbf{p}_{n-m} = \sum_{k=0}^n M_{n,m,i}^k \mathbf{p}_k.$$

Next, due to the orthogonality of polynomials, we have

$$(w_{m,i} \mathbf{p}_{n-m}, \mathbf{p}_k^\top) = \begin{cases} M_{n,m,i}^k (\mathbf{p}_k, \mathbf{p}_k^\top) & : k = n - 2m, \dots, n \\ 0 & : k < n - 2m. \end{cases} \quad (3.5)$$

The second case is trivial since $w_{m,i} \mathbf{p}_k^\top$ is at most degree $k + m \leq n - m$ and

$$(w_{m,i} \mathbf{p}_{n-m}, \mathbf{p}_k^\top) = (\mathbf{p}_{n-m}, w_{m,i} \mathbf{p}_k^\top) = 0$$

Therefore, we have

$$M_{n,m,i}^k = \begin{cases} (w_{m,i} \mathbf{p}_{n-m}, \mathbf{p}_k^\top) H_k^{-1} & : k = n - 2m, \dots, n \\ 0 & : k < n - 2m. \end{cases} \quad (3.6)$$

Furthermore, for $k = n - 2m, \dots, n$, we have

$$M_{n,m,i}^k H_k = (w_{m,i} \mathbf{p}_{n-m}, \mathbf{p}_k^\top) = (w_{m,i} \mathbf{p}_k, \mathbf{p}_{n-m}^\top)^\top = H_{n-m} (M_{n+k,m,i}^{-m})^\top.$$

Then we reindex by $k \mapsto n - m + k$ to get the result. \square

Rather than the classical three term recurrence, we now obtain a recurrence relation with $2m + 1$ terms for degree m generators. This is due to the fact that multiplication by $w_{m,i}$ increases the degree by m , and thus $(w_{m,i} \mathbf{p}_{n-m}, \mathbf{p}_k^\top)$ may be non-trivial for a larger range of degrees k , see (3.5). Here, (3.4) provides relationships between matrices for different orders of n . Two of these will be of particular importance, and we will rename them as

$$A_{n,m,i} := M_{n,m,i}^m \in \text{Mat}(r_{n-m}, r_n) \quad \text{and} \quad C_{n,m,i} := M_{n+m,m,i}^{-m} \in \text{Mat}(r_n, r_{n-m}), \quad (3.7)$$

which are related through (3.4) by

$$A_{n,m,i} H_n = H_{n-m} C_{n,m,i}^\top. \quad (3.8)$$

In particular, $A_{n,m,i}$ is the leading matrix in the recurrence relation for $w_{m,i} \mathbf{p}_{n-m}$, while $C_{n,m,i}$ is the non-leading matrix in the recurrence for $w_{m,i} \mathbf{p}_n$ which relates to $A_{n,m,i}$ via (3.8). So far, we have been working with a block orthogonal system of polynomials \mathbf{p}_n . However, if the system of polynomials is *orthonormal*, which we denote by $\tilde{\mathbf{p}}_n$, then $H_n = I$ is the identity matrix, and we obtain the following.

Corollary 3.3. Let $\tilde{\mathbf{p}}_n$ be an orthonormal system of polynomials. For $n \in \mathbb{N}_0$, $m \in [n]$, $i \in [N_m]$ and $0 \leq k \leq m$, there exist unique matrices $M_{n,m,i}^k \in \text{Mat}(r_{n-m}, r_{n-m+k})$ for such that

$$w_{m,i} \mathbf{p}_{n-m} = \sum_{k=0}^m M_{n,m,i}^k \tilde{\mathbf{p}}_{n-m+k} + \sum_{k=1}^m (M_{n-k,m,i}^k)^\top \tilde{\mathbf{p}}_{n-m-k}. \quad (3.9)$$

Furthermore, $M_{n,m,i}^{n-m}$ is symmetric.

Proof. This is immediate from Proposition 3.2 since $H_n = I$ is the identity for all n . □

3.3. Rank condition. Now, we will consider rank conditions that the matrices $A_{n,m,i}$ and $C_{n,m,i}$ must satisfy. We will consider *column joint matrices*, which stack a set of matrices in a column. For fixed $n \in \mathbb{N}$, suppose $\{Q_{n,m,i}\}_{m \in [n], i \in [N_m]}$ is a set of matrices where $Q_{n,m,i}$ is of size (p_m, q) , then we define

$$Q_{n,m} := \begin{pmatrix} Q_{n,m,1} \\ Q_{n,m,2} \\ \vdots \\ Q_{n,m,N_m} \end{pmatrix} \quad \text{and} \quad Q_n := \begin{pmatrix} Q_{n,1} \\ Q_{n,2} \\ \vdots \\ Q_{n,n} \end{pmatrix}. \quad (3.10)$$

Here $Q_{n,m} \in \text{Mat}(N_m p_m, q)$ and $Q_n \in \text{Mat}(K, q)$, where $K = \sum_{m=1}^n N_m p_m$. One particular joint matrix that will be important is $A_n \in \text{Mat}(R_n, r_n)$, built from $A_{n,m,i} \in \text{Mat}(r_{n-m}, r_n)$, where

$$R_n = \sum_{k=1}^n N_k \cdot r_{n-k}. \quad (3.11)$$

Let $G_n \in \text{Mat}(r_n, r_n)$ be the leading-coefficient matrix of \mathbf{p}_n ; that is, the matrix G_n such that²

$$\mathbf{p}_n = G_n \mathbf{w}_n + \mathbf{q}, \quad (3.12)$$

where \mathbf{w}_n consist of total degree n monomials, and \mathbf{q} has total degree strictly less than n . Next, let $L_{n,m,i} \in \text{Mat}(r_{n-m}, r_n)$ defined by

$$L_{n,m,i} \mathbf{w}_n = w_{m,i} \cdot \mathbf{w}_{n-m}.$$

Lemma 3.4. For each $n \in \mathbb{N}_0$, $m \in [n]$ and $i \in [N_m]$, the matrix $L_{n,m,i}$ satisfies

$$L_{n,m,i} \cdot L_{n,m,i}^\top = I.$$

Moreover,

$$\text{rank}(L_{n,m,i}) = r_{n-m} \quad \text{and} \quad \text{rank}(L_n) = r_n.$$

Proof. By definition, each row of $L_{n,m,i}$ has exactly one element equal to 1; otherwise the elements are 0. Thus, we have $L_{n,m,i} \cdot L_{n,m,i}^\top = I$ and $\text{rank}(L_{n,m,i}) = r_{n-m}$. Now, let

$$\mathcal{N}_n := \{\boldsymbol{\alpha} \in \mathbb{N}_0^{N_{\leq n}} : |\boldsymbol{\alpha}| = n\} \quad \text{and} \quad \mathcal{N}_{n,m,i} := \{\boldsymbol{\alpha} \in \mathcal{N}_n : \alpha_{m,i} \neq 0\},$$

where the $\boldsymbol{\alpha}$ denotes a graded exponent vector. Then, let $\mathbf{a} = (a_\alpha)_{\alpha \in \mathcal{N}_n} \in \text{Mat}(r_n, 1)$ be a column vector. Then, $L_{n,m,i}$ is a map which projects \mathbf{a} onto its restriction onto $\mathcal{N}_{n,m,i}$. Now, if $L_n \mathbf{a} = 0$, then $L_{n,m,i} \mathbf{a} = 0$ for all $m \in [n]$ and $i \in [N_m]$. However, since $\cup_{m,i} \mathcal{N}_{n,m,i} = \mathcal{N}_n$, we have $\mathbf{a} = 0$, so L_n has full rank. □

Next, for fixed $n \in \mathbb{N}$, we compare the leading coefficient matrices on each side of (3.3) to get

$$G_{n-m} L_{n,m,i} = A_{n,m,i} G_n \quad (3.13)$$

for any $m \in [n]$ and $i \in [N_m]$.

Proposition 3.5. For each $n \in \mathbb{N}$, $m \in [n]$ and $i \in [N_m]$,

$$\text{rank}(A_{n,m,i}) = \text{rank}(C_{n,m,i}) = r_{n-m}. \quad (3.14)$$

Let $A_n \in \text{Mat}(R_n, r_n)$ be the joint matrix of $A_{n,m,i}$ and $C_n^\top \in \text{Mat}(R_n, r_n)$ be the joint matrix of $C_{n,m,i}^\top \in \text{Mat}(r_{n-m}, r_n)$. Then,

$$\text{rank}(A_n) = \text{rank}(C_n^\top) = r_n. \quad (3.15)$$

²Note that these are leading coefficients in terms of the shuffle monomials (with respect to total degree). Thus, even for the monic polynomials defined by Gram-Schmidt in (2.9), G is not necessarily the identity.

Proof. From (3.13) and since all G_n are invertible, $\text{rank}(A_{n,m,i}) = r_{n-m}$ from Lemma 3.4. Furthermore, since H_n is invertible, we also have $\text{rank}(C_{n,m,i}) = r_{n-m}$ from (3.4). Now, we define

$$\mathbf{G}_m := G_m^{\oplus N_{n-m}} \in \text{Mat}(N_{n-m}r_m, N_{n-m}r_m) \quad \text{and} \quad \mathbf{G} := \mathbf{G}_{n-1} \oplus \dots \oplus \mathbf{G}_0 \in \text{Mat}(R_n, R_n). \quad (3.16)$$

Let $L_n \in \text{Mat}(R_n, r_n)$ be the joint matrix of $L_{n,m,i}$. Then, from (3.13) and the definition of the joint matrix A_n from (3.10), we have $\mathbf{G}L_n = A_n G_n$, written out in matrix form as

$$\mathbf{G}L_n = \begin{pmatrix} \mathbf{G}_{n-1} & 0 & \dots & 0 \\ 0 & \mathbf{G}_{n-2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{G}_0 \end{pmatrix} \begin{pmatrix} L_{n,1} \\ L_{n,2} \\ \vdots \\ L_{n,n} \end{pmatrix} = \begin{pmatrix} A_{n,1} \\ A_{n,2} \\ \vdots \\ A_{n,n} \end{pmatrix} G_n = A_n G_n \quad (3.17)$$

Note that \mathbf{G} is invertible and thus $\text{rank}(A_n) = \text{rank}(L_n) = r_n$. Next, we define

$$\mathbf{H}_m := H_m^{\oplus N_{n-m}} \in \text{Mat}(N_{n-m}r_m, N_{n-m}r_m) \quad \text{and} \quad \mathbf{H} := \mathbf{H}_{n-1} \oplus \dots \oplus \mathbf{H}_0 \in \text{Mat}(R_n, R_n). \quad (3.18)$$

Then, (3.8) implies $A_n H_n = \mathbf{H} C_n^T$. Because \mathbf{H} is invertible, we get $\text{rank}(C_n^T) = \text{rank}(A_n) = r_n$. \square

Because the matrix A_n has full rank, there exists a generalized left inverse $D_n^T \in \text{Mat}(r_n, R_n)$, which is expressed as the row joint matrix of $D_{n,m,i}^T \in \text{Mat}(r_n, r_{n-m})$. More explicitly,

$$D_{n,m}^T := (D_{n,m,1}^T \dots D_{n,m,N_m}^T) \quad \text{and} \quad D_n^T := (D_{n,1}^T \dots D_{n,n}^T).$$

This generalized left inverse (which may not be unique) satisfies

$$D_n^T A_n = \sum_{m=1}^n \sum_{i=1}^{N_m} D_{n,m,i}^T \cdot A_{n,m,i} = I \in \text{Mat}(r_n, r_n). \quad (3.19)$$

Proposition 3.6. *Let D_n^T be a generalized inverse of A_n . Then,*

$$\mathbf{p}_n = \sum_{m=1}^n \sum_{i=1}^{N_m} \left(w_{m,i} D_{n,m,i}^T \mathbf{p}_{n-m} - \sum_{k=-m}^{m-1} D_{n,m,i}^T M_{n,m,i}^k \mathbf{p}_{n-m+k} \right)$$

Proof. We multiply the relation from (3.3) by $D_{n,m,i}^T$ to get

$$w_{m,i} D_{n,m,i}^T \mathbf{p}_{n-m} = D_{n,m,i}^T A_{n,m,i} \mathbf{p}_n + \sum_{k=-m}^{m-1} D_{n,m,i}^T M_{n,m,i}^k \mathbf{p}_{n-m+k}.$$

Next, we sum over $m \in [n]$ and $i \in [N_m]$ to get

$$\sum_{m=1}^n \sum_{i=1}^{N_m} (w_{m,i} D_{n,m,i}^T \mathbf{p}_{n-m}) = \left(\sum_{m=1}^n \sum_{i=1}^{N_m} D_{n,m,i}^T A_{n,m,i} \right) \mathbf{p}_n + \sum_{m=1}^n \sum_{k=-m}^{m-1} \left(\sum_{i=1}^{N_m} D_{n,m,i}^T M_{n,m,i}^k \right) \mathbf{p}_{n-m+k}.$$

Then, applying (3.19), we obtain the desired result. \square

3.4. Favard's theorem. In the previous sections, we showed that given a quasi-definite functional on $\text{Sym}(W^{\text{gr}})$, orthogonal polynomials satisfy a recurrence relation where the defining matrices must satisfy certain rank conditions. Here, we will prove the converse: a generalization of Favard's theorem.

Theorem 3.7. *Let $\mathbf{p} = \{\mathbf{p}_n\}_{n=0}^\infty$ be an arbitrary sequence where $\mathbf{p}_n \in \text{Sym}(W^{\text{gr}})^{r_n}$ is a column vector of length r_n , and $\mathbf{p}_0 = 1$. Then, the following statements are equivalent.*

- 1 *There exists a quasi-definite \mathcal{L} under which $\{\mathbf{p}_n\}_{n=0}^\infty$ is a block orthogonal basis in $\text{Sym}(W^{\text{gr}})$.*
- 2 *For $n \in \mathbb{N}_0$, $m \in [n]$, $i \in [N_m]$ and $-m \leq k \leq m$, there exist $M_{n,m,i}^k \in \text{Mat}(r_{n-m}, r_{n-m+k})$ where*
 - 2.1 *the polynomials \mathbf{p}_n satisfy the relation in (3.3), and*
 - 2.2 *the matrices $A_{n,m,i} := M_{n,m,i}^m$ and $C_{n,m,i} := (M_{n+m,m,i}^{-m})^T$ satisfy (3.14) and (3.15).*

Proof. The forward direction is given by Propositions 3.2 and 3.5. We will now prove the converse direction.

\mathbf{p} is a basis. To begin, we must show that \mathbf{p} forms a basis of $\text{Sym}(W^{\text{gr}})$; it suffices to show that the leading coefficient matrix G_n of \mathbf{p}_n is invertible. Because the polynomials satisfy the relation in (3.3), the leading coefficient matrices satisfy $\mathbf{G}L_n = A_n G_n$ from (3.17) (and using the same definition of \mathbf{G} from there). We will now show that G_n is invertible by induction.

For $n = 0$, we have $\mathbf{p}_0 = 1$, so $G_0 = 1$. Now, suppose G_0, \dots, G_{n-1} are invertible. By the definition of \mathbf{G} in (3.16), this implies that \mathbf{G} is invertible. Therefore, we have

$$\text{rank}(A_n G_n) = \text{rank}(\mathbf{G} L_n) = \text{rank}(L_n) = r_n,$$

by Lemma 3.4. Then, using the rank hypothesis of $\text{rank}(A_n) = r_n$, and the rank inequality for product matrices, we have

$$\text{rank}(G_n) \geq \text{rank}(A_n G_n) \geq \text{rank}(A_n) + \text{rank}(G_n) - r_n = \text{rank}(G_n).$$

Thus, $\text{rank}(G_n) = \text{rank}(A_n G_n) = r_n$, so G_n is invertible.

p is block orthogonal. Because \mathbf{p} is a basis of $\text{Sym}(W^{\text{gr}})$, we can define $\mathcal{L} : \text{Sym}(W^{\text{gr}}) \rightarrow \mathbb{R}$ by

$$\mathcal{L}(1) = 1, \quad \mathcal{L}(\mathbf{p}_n) = 0 \text{ for } n \geq 1.$$

Now, we will show that \mathbf{p} is block orthogonal with respect to this linear functional. In particular, we will use induction to show that

$$\mathcal{L}(\mathbf{p}_k \mathbf{p}_j^T) = 0 \text{ for } k \neq j. \tag{3.20}$$

Suppose (3.20) holds for k, j such that $0 \leq k \leq n-1$ and $j > k$. Note that this holds by definition when $n = 1$. Now, we note that Proposition 3.6 only used the recurrence in (3.3) and the rank condition in (3.15). Thus, for $\ell > n$, we have

$$\mathcal{L}(\mathbf{p}_n \mathbf{p}_\ell^T) = \mathcal{L} \left(\sum_{m=1}^n \sum_{i=1}^{N_m} w_{m,i} D_{n,m,i}^T \mathbf{p}_{n-m} \mathbf{p}_\ell^T \right) = \mathcal{L} \left(\sum_{m=1}^n \sum_{i=1}^{N_m} D_{n,m,i}^T \mathbf{p}_{n-m} \left(\sum_{k=-m}^m M_{\ell+m,m,i}^k \mathbf{p}_{\ell+k} \right)^T \right) = 0,$$

where we use Proposition 3.6 and the induction hypothesis in the first equality, the relation in (3.3) on $w_{m,i} \mathbf{p}_\ell^T$ in the second, and the induction hypothesis in the third. Thus, (3.20) holds.

L is nondegenerate. Now, we will show that $H_n = \mathcal{L}(\mathbf{p}_n \mathbf{p}_n^T)$ is invertible. First, we note that H_n is symmetric by definition. Next, by block orthogonality, (3.6) holds, and therefore (3.8) also holds. Now, using the definition of \mathbf{H} from (3.18), we have

$$A_n H_n = \mathbf{H} C_n^T. \tag{3.21}$$

We will once again show that H_n is invertible by induction. Because $\mathcal{L}(1) = 1$, we have $H_0 = 1$, which is invertible. Now, suppose that H_0, \dots, H_{n-1} is invertible, which implies that \mathbf{H} is invertible. Then, using the rank condition $\text{rank}(C_n^T) = r_n$ from (3.15) by hypothesis, we have

$$\text{rank}(A_n H_n) = \text{rank}(\mathbf{H} C_n^T) = \text{rank}(C_n^T) = r_n$$

Then, using the rank inequality for product matrices and the rank condition for A_n again,

$$\text{rank}(H_n) \geq \text{rank}(A_n H_n) \geq \text{rank}(A_n) + \text{rank}(H_n) - r_n = \text{rank}(H_n).$$

Thus, $\text{rank}(H_n) = \text{rank}(A_n H_n) = r_n$, which implies that H_n is invertible.

Finally, the fact that H_n is invertible implies that \mathcal{L} is a quasi-definite linear functional which makes \mathbf{p} a block orthogonal basis of $\text{Sym}(W^{\text{gr}})$. □

Next, we will show Favard's theorem for positive-definite functionals.

Theorem 3.8. Let $\mathbf{p} = \{\tilde{\mathbf{p}}_n\}_{n=0}^\infty$ be an arbitrary sequence where $\tilde{\mathbf{p}}_n \in \text{Sym}(W^{\text{gr}})^{r_n}$ is a column vector of length r_n , and $\tilde{\mathbf{p}}_0 = 1$. Then, the following statements are equivalent.

- 1 There exists a positive-definite \mathcal{L} under which $\{\tilde{\mathbf{p}}_n\}_{n=0}^\infty$ is an orthonormal basis in $\text{Sym}(W^{\text{gr}})$.
- 2 For $n \in \mathbb{N}_0$, $m \in [n]$, $i \in [N_m]$ and $0 \leq k \leq m$, there exist matrices $M_{n,m,i}^k \in \text{Mat}(r_{n-m}, r_{n-m+k})$ such that
 - 2.1 the polynomials $\tilde{\mathbf{p}}_n$ satisfy the relation in (3.9), and
 - 2.2 the matrices $A_{n,m,i} := M_{n,m,i}^m$ and $C_{n,m,i} := (M_{n-m,m,i}^m)^T$ satisfy (3.14) and (3.15).

Proof. The forward direction is given by Corollary 3.3 and Proposition 3.5. We will now prove the converse direction. From Theorem 3.7, there exists a quasi-definite linear functional \mathcal{L} such that $\tilde{\mathbf{p}}_n$ is a block orthogonal basis in $\text{Sym}(W^{\text{gr}})$, so we only need to show that \mathcal{L} is positive-definite. Let $H_n = \mathcal{L}(\tilde{\mathbf{p}}_n \tilde{\mathbf{p}}_n^T)$; it suffices to show that this is the identity for every $n \in \mathbb{N}_0$. Because $\tilde{\mathbf{p}}_0 = 1$ and $\mathcal{L}(1) = 1$, we have $H_0 = 1$, and we proceed by induction on n . Suppose $H_k = I$ is the identity for $k \leq n-1$. Following the previous proof, $C_n = (A_n)^T$ implies that (3.21) becomes

$$A_n H_n = \mathbf{H} A_n.$$

By definition of \mathbf{H} from (3.18) and induction, we have $\mathbf{H} = I$, so $H_n = I$ is the identity. □

3.5. Jacobi matrices and measures on \widehat{W} . Favard's theorem in Theorem 3.8 only provides the existence of a positive-definite linear functional on $\text{Sym}(W^{\text{gr}})$. In this section, we consider the question of when the functional \mathcal{L} is induced by a measure $\mu \in \mathcal{M}(\widehat{W})$ on $\widehat{W} = \prod_{m=0}^{\infty} W^m$,

$$\mathcal{L}(P) = \int_{\widehat{W}} P(w) d\mu(w). \quad (3.22)$$

In particular, for the remainder of this section, we will work exclusively with positive-definite linear functionals, and the corresponding recurrence matrix conditions in Theorem 3.8. We do this by generalizing the notion of *Jacobi matrices*, and use the spectral theory of commuting self adjoint operators (CSOs). In this section, we will be exclusively working with *orthonormal polynomials*, and thus use the recurrence relation in Corollary 3.3.

Suppose we have a positive-definite linear functional \mathcal{L} on $\text{Sym}(W^{\text{gr}})$, and let $\widetilde{\mathbf{p}} = (\widetilde{\mathbf{p}}_n)_{n=0}^{\infty}$ denote a system of orthonormal polynomials with respect to this functional. Furthermore, let $H(W^{\text{gr}}) \cong \ell^2$ denote the Hilbert space completion of $\text{Sym}(W^{\text{gr}})$, equipped with an orthonormal basis³ denoted ϕ_{α} as in (3.2). We let $\Phi_n = (\phi_{\alpha})_{|\alpha|=n}$ denote the column vector of total degree n elements. Then, for each $m \in \mathbb{N}$ and $i \in [N_m]$, we collect all recurrence matrices $M_{n,m,i}^k$ into the *Jacobi matrix* $J_{m,i}$, which is a linear operator in sequence space $H(W^{\text{gr}})$. This is defined as a semi-infinite band-diagonal block matrix, where the width of the band is $2m + 1$, as

$$J_{m,i} := \begin{pmatrix} M_{m,m,i}^0 & M_{m,m,i}^1 & M_{m,m,i}^2 & \cdots & M_{m,m,i}^m & 0 & 0 & \cdots \\ (M_{m,m,i}^1)^{\top} & M_{m+1,m,i}^0 & M_{m+1,m,i}^1 & \cdots & M_{m+1,m,i}^{m-1} & M_{m+1,m,i}^m & 0 & \cdots \\ (M_{m,m,i}^2)^{\top} & (M_{m+1,m,i}^1)^{\top} & M_{m+2,m,i}^0 & \cdots & M_{m+2,m,i}^{m-2} & M_{m+2,m,i}^{m-1} & M_{m+2,m,i}^m & \cdots \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \quad (3.23)$$

where the $M_{n,m,i}^k$ are given in Corollary 3.3. Note that the diagonal blocks are symmetric matrices. Now, our aim is to show there exists a probability measure μ on \widehat{W} such that $J_{m,i}$ is unitarily equivalent to the multiplication by $w_{m,i}$ operator on $L^2(\widehat{W}, \mu)$.

3.5.1. Background on spectral theory. We do this by using the spectral theory of a countable family of CSOs. The following summary of the relevant theory is from [52]. Suppose \mathcal{H} is a separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$. Every self-adjoint operator $T : \mathcal{H} \rightarrow \mathcal{H}$ has a spectral measure E on \mathbb{R} . In particular, E is a projection-valued Borel measure such that $T = \int_{\mathbb{R}} \lambda dE(\lambda)$, $E(\mathbb{R})$ is the identity operator and $E(B \cap C) = E(B) \cap E(C)$. We say that a family $\{T_j\}_{j=1}^{\infty}$ of self-adjoint operators *strongly commute* if their spectral measures commute,

$$E_i(B)E_j(C) = E_j(C)E_i(B)$$

for any Borel sets $B, C \subset \mathbb{R}$. If a finite family $\{T_j\}_{j=1}^d$ commute, then the *spectral measure of the commuting family* T_1, \dots, T_d is a projection valued measure on \mathbb{R}^d defined by

$$E(B_1 \times \dots \times B_d) = E_1(B_1) \cdots E_d(B_d).$$

If we have a countably infinite family $\{T_j\}_{j=1}^{\infty}$, we want to obtain a spectral measure on the vector space of sequences \mathbb{R}^{∞} equipped with the product topology. Let $\mathcal{B}(\mathbb{R}^{\infty})$ denote its Borel σ -algebra.

Definition 3.9. [52, Definition 1] An operator-valued measure E defined on $(\mathbb{R}^{\infty}, \mathcal{B}(\mathbb{R}^{\infty}))$ is a *spectral measure* (or a *resolution of the identity*) if:

- 1 **(projection)** $E(B)$ is a projection on \mathcal{H} for all $B \in \mathcal{B}(\mathbb{R}^{\infty})$, $E(\emptyset) = 0$ and $E(\mathbb{R}^{\infty}) = I$;
- 2 **(additivity)** if $\{B_j\}_{j=1}^{\infty}$ are mutually disjoint, then $E(\bigcup_{j=1}^{\infty} B_j) = \sum_{j=1}^{\infty} E(B_j)$;
- 3 **(orthogonality)** for all $B, B' \in \mathcal{B}(\mathbb{R}^{\infty})$, we have $E(B \cap B') = E(B)E(B')$.

Theorem 3.10. [52, Theorem 1] For every countable family $\{T_j\}_{j=1}^{\infty}$ of CSOs, there exists a unique spectral measure E . Conversely, every spectral measure is generated by a countable family of CSOs where

$$T_j = \int_{\mathbb{R}^{\infty}} \lambda_j dE(\lambda_1, \lambda_2, \dots) = \int_{\mathbb{R}} \lambda_j dE_j(\lambda_j), \quad (3.24)$$

where E_j is the 1D spectral measure $E_j(B) = E(\mathbb{R} \times \dots \times \mathbb{R} \times B \times \mathbb{R} \times \dots)$, where B is in the j^{th} spot.

For a countable family $\{T_j\}_{j=1}^{\infty}$, a *cyclic vector* is an element $\Phi_0 \in \mathcal{H}$ such that

$$\overline{\text{span}\{E(B)\Phi_0 : B \in \mathcal{B}(\mathbb{R}^{\infty})\}} = \mathcal{H}. \quad (3.25)$$

The following spectral theorem is the main result we will need.

³We treat $H(W^{\text{gr}})$ as an abstract Hilbert space, and use the notation ϕ_{α} to forget that these elements are polynomials.

Theorem 3.11. [52, Theorem 4] Let $\{T_j\}_{j=1}^\infty$ be a countable family of CSOs on \mathcal{H} , equipped with a cyclic vector $\Phi_0 \in \mathcal{H}$ (in the sense of (3.25)). Then, there exists a probability measure $\mu \in \mathcal{M}(\mathbb{R}^\infty)$ and a unitary transformation $U : \mathcal{H} \rightarrow L^2(\mathbb{R}^\infty, \mu)$ such that

$$T_j = U^{-1} \lambda_j U, \tag{3.26}$$

where $\lambda_j : L^2(\mathbb{R}^\infty, \mu) \rightarrow L^2(\mathbb{R}^\infty, \mu)$ is the multiplication operator by the j^{th} independent variable.

To apply this, we must show that the $J_{m,i}$ are CSOs and the existence of a cyclic vector.

3.5.2. *Commutativity and cyclic vectors.* Now, we return to the Jacobi matrices and start by showing conditions for commutativity along with establishing the existence of a cyclic vector.

Proposition 3.12. Let $\tilde{\mathbf{p}}$ be an orthonormal basis which satisfies the relations in Corollary 3.3. Then for all $n_1 \leq n_2 \in \mathbb{N}$, $m_1, m_2 \in \mathbb{N}$ and $i \in [N_{m_1}], j \in [N_{m_2}]$, we have

$$\begin{aligned} & \sum_{r \in R_{1,1}^1} (M_{m_1+r, m_1, i}^{n_1-r})^T M_{m_2+r, m_2, j}^{n_2-r} + \sum_{r \in R_{1,2}^2} M_{n_1+m_1, m_1, i}^{r-n_1} M_{m_2+r, m_2, j}^{n_2-r} + \sum_{r \in R_{1,2}^3} M_{n_1+m_1, m_1, i}^{r-n_1} (M_{n_2+m_2, m_2, j}^{r-n_2})^T \\ &= \sum_{r \in R_{2,1}^1} (M_{m_2+r, m_2, j}^{n_1-r})^T M_{m_1+r, m_1, i}^{n_2-r} + \sum_{r \in R_{2,1}^2} M_{n_1+m_2, m_2, j}^{r-n_1} M_{m_1+r, m_1, i}^{n_2-r} + \sum_{r \in R_{2,1}^3} M_{n_1+m_2, m_2, j}^{r-n_1} (M_{n_2+m_1, m_1, i}^{r-n_2})^T, \end{aligned} \tag{3.27}$$

where

$$R_{i,j}^1 = [n_1 - m_i, n_1] \cap [n_2 - m_j, n_2], \tag{3.28}$$

$$R_{i,j}^2 = [n_1 + 1, n_1 + m_i] \cap [n_2 - m_j, n_2], \tag{3.29}$$

$$R_{i,j}^3 = [n_1 + 1, n_1 + m_i] \cap [n_2 + 1, n_2 + m_j]. \tag{3.30}$$

Proof. We obtain these relations by applying the relations in Corollary 3.3 to the equality

$$(w_{m_1, i} \tilde{\mathbf{p}}_{n_1}, w_{m_2, j} \tilde{\mathbf{p}}_{n_2}) = (w_{m_2, j} \tilde{\mathbf{p}}_{n_1}, w_{m_1, i} \tilde{\mathbf{p}}_{n_2}). \tag{3.31}$$

□

Lemma 3.13. The basis vector $\Phi_0 = \phi_0 \in H(W^{\text{gr}})$ is a cyclic vector for $\{J_{m,i}\}$.

Proof. Let $H_E = \overline{\text{span}\{E(B)\Phi_0 : B \in \mathcal{B}(\widehat{W})\}} \subset H(V)$. Then by the spectral description of $J_{m,i}$ in Theorem 3.10, we have $E(B)J_{m,i} = J_{m,i}E(B)$ for any $B \in \mathcal{B}(\widehat{W})$, so that $J_{m,i}H_E \subset H_E$. Next, by definition of the Jacobi matrices, we have

$$J_{m,i}\Phi_{n-m} = \sum_{k=0}^m M_{n,m,i}^k \Phi_{n-m+k} + \sum_{k=1}^m (M_{n-k, m, i}^k)^T \Phi_{n-m-k}, \tag{3.32}$$

and thus, by the same arguments as Proposition 3.6, we have

$$\Phi_n = \sum_{m=1}^n \sum_{i=1}^{N_m} (J_{m,i} D_{n,m,i}^T \Phi_{n-m}) + \sum_{k=0}^{n-1} E_n^k \Phi_k.$$

Thus, this provides a polynomial in P_n in the $J_{m,i}$ such that $\Phi_n = P_n(J_{m,i})\Phi_0$. Then, since the Φ_n are dense in $H(W^{\text{gr}})$, we have $H_E = H(W^{\text{gr}})$. □

3.5.3. *Bounded case.* Next, we move on to showing that the Jacobi matrices $J_{m,i}$ are a commuting self-adjoint family of operators on $H(W^{\text{gr}})$. The arguments depend on whether we assume that $J_{m,i}$ is bounded, and we begin with the bounded setting. These results generalize the classical setting, and while we detail aspects of the proof which differ in our setting, we will defer parts of the proof to existing references in situations where the proof is the same.

Lemma 3.14. The operator $J_{m,i}$ is bounded if and only if $\sup_{n \geq 0} \|M_{n,m,i}^k\|_2 < \infty$ for all $k \in [-m, m]$.

Proof. The proof is analogous to [25, Lemma 3.4.3]. □

Theorem 3.15. Let $\tilde{\mathbf{p}} = (\tilde{\mathbf{p}}_n)_{n=0}^\infty$ be a sequence in $\text{Sym}(W^{\text{gr}})$ such that $\tilde{\mathbf{p}}_0 = 1$. The following are equivalent.

- 1 There exists a determinate measure $\mu \in \mathcal{M}(\widehat{W})$ with compact support such that $\tilde{\mathbf{p}}$ is orthonormal with respect to μ .
- 2 Statement (2) in Theorem 3.8 holds and in addition, $\sup_{n \geq 0} \|M_{n,m,i}^k\|_2 < \infty$ for all $k \in [-m, m]$.

Proof. First, if μ has compact support, then the multiplication operators $T_{w_{m,i}} : L^2(\widehat{W}, \mu) \rightarrow L^2(\widehat{W}, \mu)$ are bounded. Because these multiplication operators are represented by the Jacobi matrices $J_{m,i}$ in the $\widetilde{\mathbf{p}}_n$ basis, the result follows from Lemma 3.14.

Next, suppose the second statement holds, so by Theorem 3.8, $\widetilde{\mathbf{p}}$ are orthonormal with respect to a positive-definite linear functional \mathcal{L} . We define the Jacobi matrices as in (3.23), which are bounded by Lemma 3.14. In the bounded setting, strong commutativity is equivalent to commutativity of operators. We can directly verify that the conditions in Proposition 3.12 are equivalent to $J_{m_1,i} J_{m_2,j} = J_{m_2,j} J_{m_1,i}$; thus the $J_{m,i}$ are strongly commutative. Next, it is clear by definition of the $J_{m,i}$ that they are symmetric, and since they are bounded, they must be self-adjoint. Because there exists a cyclic vector (Lemma 3.13), we apply Theorem 3.11 to obtain a measure $L^2(\widehat{W}, \mu)$ such that $J_{m,i}$ are unitarily equivalent to multiplication operators $T_{m,i}$, so that

$$\int_{\widehat{W}} \widetilde{\mathbf{p}}_n(w) \widetilde{\mathbf{p}}_m^T(w) d\mu(w) = \langle \widetilde{\mathbf{p}}_n(T) \mathbf{1}, \widetilde{\mathbf{p}}_m^T(T) \mathbf{1} \rangle_{L^2(\widehat{W}, \mu)} = \langle \widetilde{\mathbf{p}}_n(J) \Phi_0, \widetilde{\mathbf{p}}_m^T(J) \Phi_0 \rangle_{H(W^{\text{gr}})} = \langle \Phi_n, \Phi_m^T \rangle_{H(W^{\text{gr}})}.$$

Thus, $\widetilde{\mathbf{p}}$ is orthonormal with respect to μ . The support $S_{m,i}$ of the spectral measure $E_{m,i}$ of $J_{m,i}$ is compact since $J_{m,i}$ is bounded. Thus, the support of $S = \prod_{m,i} S_{m,i}$ is compact by Tychonoff's theorem. Finally, we note that the projection of μ to any finite dimensional subspace is determinate (since it is compact), and therefore, μ is determinate by [2, Corollary 5.3]. \square

3.5.4. Unbounded case. Next, we consider the case where the Jacobi matrices $J_{m,i}$ are unbounded operators $J_{m,i} : \mathcal{D}(J_{m,i}) \rightarrow H(W^{\text{gr}})$, where the domain $\mathcal{D}(J_{m,i}) \subset H(W^{\text{gr}})$ consists of all sequences $f \in H(W^{\text{gr}})$ such that $J_{m,i} f \in H(W^{\text{gr}})$.

Lemma 3.16. *Suppose*

$$\sum_{n=0}^{\infty} \frac{1}{\mathfrak{M}_{n,m,i}} = \infty \quad \text{where} \quad \mathfrak{M}_{n,m,i} = \sum_{k=1}^m \sum_{r=n+1}^{n+k} \|M_{m+r-k,m,i}^k\|_2. \quad (3.33)$$

Then, $J_{m,i}$ is self-adjoint.

Proof. Let $f, g \in \mathcal{D}(J_{m,i})$ such that $f = \sum \mathbf{a}_k^T \Phi_k$ and $g = \sum \mathbf{b}_k^T \Phi_k$. We define $\mathbf{a}_k = \mathbf{b}_k = 0$ for $k < 0$. We will prove that $J_{m,i}$ is symmetric so that $\langle J_{m,i} f, g \rangle = \langle J_{m,i} g, f \rangle$. By definition of the $J_{m,i}$, we have

$$\langle J_{m,i} f, g \rangle = \sum_{r=0}^{\infty} \left(\sum_{k=1}^m \mathbf{a}_{r-k}^T M_{m+r-k,m,i}^k \mathbf{b}_r + \mathbf{a}_r^T M_{m+r,m,i}^0 \mathbf{b}_r + \sum_{k=1}^m \mathbf{a}_{r+k}^T (M_{m+r,m,i}^k)^T \mathbf{b}_r \right). \quad (3.34)$$

For $n \in \mathbb{N}$, we define the truncation by

$$S_n(\langle J_{m,i} f, g \rangle) = \sum_{r=0}^n \left(\sum_{k=1}^m \mathbf{a}_{r-k}^T M_{m+r-k,m,i}^k \mathbf{b}_r + \mathbf{a}_r^T M_{m+r,m,i}^0 \mathbf{b}_r + \sum_{k=1}^m \mathbf{a}_{r+k}^T (M_{m+r,m,i}^k)^T \mathbf{b}_r \right), \quad (3.35)$$

and note that $\langle J_{m,i} f, g \rangle = \lim_{n \rightarrow \infty} S_n(\langle J_{m,i} f, g \rangle)$. Then, by direct computation using that $M_{m+r-k,m,i}^k$ is symmetric, we have

$$S_n(\langle J_{m,i} f, g \rangle) - S_n(\langle J_{m,i} g, f \rangle) = \sum_{k=1}^m \sum_{r=n+1}^{n+k} -\mathbf{a}_{r-k}^T M_{m+r-k,m,i}^k \mathbf{b}_r + \mathbf{b}_{r-k}^T M_{m+r-k,m,i}^k \mathbf{a}_r. \quad (3.36)$$

Then, we can bound this by

$$\begin{aligned} |S_n(\langle J_{m,i} f, g \rangle) - S_n(\langle J_{m,i} g, f \rangle)| &\leq \sum_{k=1}^m \sum_{r=n+1}^{n+k} \|M_{m+r-k,m,i}^k\|_2 (\|\mathbf{a}_{r-k}\|_2^2 + \|\mathbf{b}_r\|_2^2 + \|\mathbf{b}_{r-k}\|_2^2 + \|\mathbf{a}_r\|_2^2) / 2 \\ &\leq \sum_{k=1}^m \sum_{r=n+1}^{n+k} \|M_{m+r-k,m,i}^k\|_2 \left(\sum_{s=n-m+1}^{n+m} \|\mathbf{a}_s\|_2^2 + \|\mathbf{b}_s\|_2^2 \right). \end{aligned}$$

Then, suppose $|\langle J_{m,i} f, g \rangle - \langle J_{m,i} g, f \rangle| = \delta$. Then, for sufficiently large N , we have $|S_n(\langle J_{m,i} f, g \rangle) - S_n(\langle J_{m,i} g, f \rangle)| > \delta/2$ for all $n \geq N$, and therefore

$$\frac{\delta}{2} \sum_{n \geq N} \frac{1}{\mathfrak{M}_{n,m,i}} \leq \sum_{n \geq N} \sum_{s=n-m+1}^{n+m} \|\mathbf{a}_s\|_2^2 + \|\mathbf{b}_s\|_2^2 \leq 2m(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2) \leq 2m(\|f\|^2 + \|g\|^2) < \infty, \quad (3.37)$$

so by the hypothesis, $\delta = 0$ and $J_{m,i}$ is symmetric. Then, the fact that $J_{m,i}$ is self-adjoint can be shown in the same way as [61, Lemma 1]. \square

Lemma 3.17. *Suppose (3.33) holds. Then the operators $J_{m,i}$ pairwise strongly commute.*

Proof. The proof is analogous to [61, Lemma 3] by applying [45, Lemma 9.2] to the matrices $J_{m,i}$. In particular, we know from Proposition 3.12 that the $J_{m,i}$ commute on the dense subspace $\mathcal{D} = \text{Sym}(W^{\text{gr}}) \subset H(W^{\text{gr}})$. \square

Theorem 3.18. *Suppose $\tilde{\mathbf{p}} = (\tilde{\mathbf{p}}_n)_{n=0}^\infty$ is a sequence in $\text{Sym}(W^{\text{gr}})$ such that $\tilde{\mathbf{p}}_0 = 1$, statement (2) in Theorem 3.8 holds, and (3.33) holds. Then, there exists a determinate measure $\mu \in \mathcal{M}(\widehat{W})$ such that $\tilde{\mathbf{p}}$ is orthonormal with respect to μ .*

Proof. The proof is the same as Theorem 3.15 aside from determinacy, where Lemmas 3.16 and 3.17 are used to show that the $J_{m,i}$ are commuting self-adjoint operators. Thus, there exists a measure $\mu \in \mathcal{M}(\widehat{W})$ such that $\tilde{\mathbf{p}}$ is orthonormal with respect to μ .

In order to prove determinacy, consider a finite dimensional subspace $U \subset W$ which is defined by a finite subset of the basis vectors $\{w_{m_k,i_k}\}_{k=0}^K$ of W . Let $\mu_U \in \mathcal{M}(U)$ denote the pushforward of μ under the projection to U . The induced inner product defined on $\text{Sym}(U^{\text{gr}})$ is simply the restriction of the inner product in $\text{Sym}(W^{\text{gr}})$. Let $\tilde{\mathbf{p}}^U$ be the collection of orthonormal polynomials in $\tilde{\mathbf{p}}$ which are valued in U . Furthermore, both Lemmas 3.16 and 3.17 hold for $\tilde{\mathbf{p}}^U$. Then, by the same arguments as above, we have a measure $\mu'_U \in \mathcal{M}(U)$ such that $\tilde{\mathbf{p}}^U$ is orthonormal with respect to μ'_U . In particular, all multiplication operators corresponding⁴ to $\{J_{m_k,i_k}^U\}_{k=0}^K$ are self-adjoint, so by [54, Theorem 14.2], μ'_U is determinate, so $\mu_U = \mu'_U$. Finally, this holds for any finite dimensional $U \subset W$ constructed from a finite set $\{w_{m_k,i_k}\}_{k=0}^K$ of basis elements, μ is determinate by [2, Corollary 5.3]. \square

3.6. Measures on $G(V)$. Now, we return to our original setting of shuffle polynomials and set $W = L(V)$ so that $\text{Sym}(W^{\text{gr}}) \cong \text{Sh}(V^*)$ as commutative algebras. Recall that in order to recover our interpretation of $\text{Sh}(V^*)$ as signature polynomials, we evaluate polynomials $p \in \text{Sh}(V^*)$ on the signature of paths. The positive definite Favard's theorem in Theorem 3.8 immediately holds in this setting as it only specifies an inner product on $\text{Sh}(V^*)$. However, when we consider Theorems 3.15 and 3.18, which shows the existence of a measure μ on \widehat{W} , this leads to a distinction which is absent in the classical setting: the measure μ may not come from a measure on path space.

First, $\widehat{L}(V)$ is equipped with the product topology and consists of free Lie series on V [50]. Then, when $G(V)$ is equipped with the projective topology, the tensor exponential $\exp_\otimes : \widehat{L}(V) \rightarrow G(V)$ is a homeomorphism. Note that for any $p \in \text{Sh}(V^*)$ and $w \in \widehat{L}(V)$, we have

$$p(w) = \langle p, \exp_\otimes(w) \rangle. \tag{3.38}$$

Thus, given a measure $\mu \in \mathcal{M}(\widehat{L}(V))$, the pushforward $\mu_G = (\exp_\otimes)_* \mu \in \mathcal{M}(G(V))$ satisfies

$$\int_{\widehat{L}(V)} p(w)q(w)d\mu(w) = \int_W \langle p \sqcup q, \exp_\otimes(w) \rangle d\mu(w) = \int_{G(V)} \langle p \sqcup q, x \rangle d\mu_G(x). \tag{3.39}$$

Thus, inner products on $\text{Sym}(L(V)^{\text{gr}})$ induced by measures on $\widehat{L}(V)$ correspond to inner products on $\text{Sh}(V^*)$ induced by measures on $G(V)$. However, this need not imply μ_G is induced by a measure on path space. In fact, we have the following inclusions of groups

$$\text{im}(S) \subset \widetilde{G}(V) \subset G(V), \tag{3.40}$$

where S denotes the signature, so we must further consider the support of μ_G . While the signature of paths have an infinite radius of convergence, so $\text{im}(S) \subset \widetilde{G}(V)$, the characterization of the image is an open problem [33, Problem 1.12]. Therefore, we restrict our attention to the first step and consider when the measure μ_G is supported in $\widetilde{G}(V)$. We start with the bounded case of Theorem 3.15.

Proposition 3.19. *Suppose $\tilde{\mathbf{p}} = (\tilde{\mathbf{p}}_n)_{n=0}^\infty$ be a sequence in $\text{Sym}(L(V)^{\text{gr}})$ such that $\tilde{\mathbf{p}}_0 = 1$. Suppose statement (2) in Theorem 3.8 holds, and in addition the Jacobi matrices $J_{m,i}$ are bounded and satisfy*

$$\sum_{m,i} \|J_{m,i}\| \lambda^m < \infty \text{ for all } \lambda > 0. \tag{3.41}$$

Then, there exists a determinate measure $\tilde{\mu}_G \in \mathcal{M}(\widetilde{G}(V))$ such that $\tilde{\mathbf{p}}$ is orthonormal with respect to $\tilde{\mu}_G$.

Proof. Because the Jacobi matrices are bounded, Lemma 3.14 implies that statement (1) of Theorem 3.15 holds, so there exists a compact measure $\mu \in \mathcal{M}(\widehat{W})$. Recall that this measure is defined by $\mu(B) = \langle E(B)\Phi_0, \Phi_0 \rangle_{H(L(V))}$, where E is the spectral measure with respect to the family $\{J_{m,i}\}$ of Jacobi matrices. Let $E_{m,i}$ denote the 1D spectral measure of the operator $J_{m,i}$, which is supported on the spectrum $\sigma(J_{m,i})$. As $J_{m,i}$ is a bounded, its spectral radius is bounded by $\|J_{m,i}\|$.

⁴Note that the matrices $M_{n,m,i}^k$ in the recurrence relation for $\tilde{\mathbf{p}}^U$ are different in general. We let J_{m_k,i_k}^U denote the Jacobi operators for $\tilde{\mathbf{p}}^U$.

By (3.41), this implies that the support of the joint spectral measure μ on W consists of elements $\mathbf{a} = \sum_m \mathbf{c}_m^T \mathbf{w}_m$, where \mathbf{w}_m denotes the column vector of degree m basis elements of W , such that

$$\sum_m \|\mathbf{c}_m\| \lambda^m < \infty \quad (3.42)$$

for all $\lambda > 0$. Let $\exp_{\otimes}^n(\mathbf{a})$ denote the degree n component of $\exp_{\otimes}(\mathbf{a})$. This implies that

$$\sum_{n=0}^{\infty} \|\exp_{\otimes}^n(\mathbf{a})\| \lambda^n \leq \sum_{n=0}^{\infty} \exp\left(\sum_{m=0}^{\infty} \lambda^m \|\mathbf{c}_m\|\right) < \infty, \quad (3.43)$$

where we assume that the Lie algebra basis $\ell_{m,i} = \iota(\mathbf{w}_{m,i})$ are normalized so that $\|\ell_{m,i}\|_{T(V)} = 1$ for all m, i . Thus $\tilde{\mu} = \beta_* \mu \in \mathcal{M}(\tilde{G}(V))$. This also implies that the expectation of $\tilde{\mu}_G$ will have infinite radius of convergence, so $\tilde{\mu}_G$ is determinate by [19, Theorem 6.1]. \square

This condition on the Jacobi matrices imply that the support of $\tilde{\mu}$ consist of group-like elements whose logarithm (viewed as elements of $T(V)$) itself has infinite radius of convergence. However, a conjecture by Lyons-Sidarova [44], and recently refined in [12] states that: the log-signature of a tree-reduced bounded variation path has infinite radius of convergence if and only if it is conjugate to a line segment. If true, this conjecture places extremely strong conditions on a potential measure on path space which induces $\tilde{\mu}$.

In order to obtain infinite radius of convergence in the case of unbounded Jacobi operators, we consider a truncated setting where we restrict $L(V)$ to Lie elements of degree at most M , denoted $L_M(V)$. In this case, we take $W = L_M(V)$, and we define

$$G_M(V) = \{\beta(\mathbf{w}) : \mathbf{w} \in L_M(V)\} \subset \tilde{G}(V). \quad (3.44)$$

Note that both $\text{Sym}(L_M(V)^{\text{gr}})$ and $G_M(V)$ still contain elements of arbitrary degree (in particular, $\text{Sym}(L_M(V)^{\text{gr}})$ is still infinite dimensional).

Proposition 3.20. *Suppose $\tilde{\mathbf{p}} = (\tilde{\mathbf{p}}_n)_{n=0}^{\infty}$ is a sequence in $\text{Sym}(L_M(V)^{\text{gr}})$ such that $\tilde{\mathbf{p}}_0 = 1$. Suppose statement (2) in Theorem 3.8 holds, and the matrices satisfy (3.33). Then, there exists a determinate measure $\tilde{\mu}_G \in \mathcal{M}(G_M(V))$ such that $\tilde{\mathbf{p}}$ is orthonormal with respect to $\tilde{\mu}_G$.*

Proof. Applying Theorem 3.18 to $W = L_M(V)$, there exists a measure $\mu \in \mathcal{M}(L_M(V))$ such that $\tilde{\mathbf{p}}$ is orthonormal with respect to its induced linear functional. Then, the pushforward by the restricted tensor exponential $\exp_{\otimes} : L_M(V) \rightarrow G_M(V)$ yields a measure $\mu_G \in \mathcal{M}(G_M(V))$. As the expectation of μ_G has infinite radius of convergence in $G(V)$, the measure μ_G is determinate by [19, Theorem 6.1]. \square

In fact, when we further truncate the total degree of polynomials considered, we can take advantage of Chow's theorem [21], which shows the surjectivity of the truncated path signature. This leads to a measure on path space which induces the inner product up to polynomials of the truncated total order.

Theorem 3.21. *Suppose $\tilde{\mathbf{p}} = (\tilde{\mathbf{p}}_n)_{n=0}^{\infty}$ is a sequence in $\text{Sym}(L_M(V)^{\text{gr}})$ such that $\tilde{\mathbf{p}}_0 = 1$. Suppose statement (2) in Theorem 3.8 holds, and the matrices satisfy (3.33). Then, there exists a measure $\rho_M \in \mathcal{M}(C^{1-\text{var}}([0, T], V))$ such that $(\tilde{\mathbf{p}}_n)_{n=0}^M$ is orthonormal with respect to $S_* \rho_M \in \mathcal{M}(\tilde{G}(V))$.*

Proof. Applying Theorem 3.18 to $W = L_M(V)$, there exists a measure $\mu \in \mathcal{M}(L_M(V))$ such that $\tilde{\mathbf{p}}$ is orthonormal with respect to its induced linear functional. Then, by the generalized Tchakaloff's theorem [8, Theorem 2], there exists a cubature measure $\mu_M = \sum_{i=1}^N \lambda_i \delta_{\ell_i} \in \mathcal{M}(L_M(V))$ of order M with respect to generators with inhomogeneous grading, where $N \in \mathbb{N}$, $\lambda_i \in \mathbb{R}$, and $\ell_i \in L_M(V)$. In particular, $(\tilde{\mathbf{p}}_n)_{n=0}^M$ is orthonormal with respect to μ_M . Then, by Chow's theorem [21] (see also [30, Theorem 7.28]), there exists a piecewise linear path $X_i \in C^{1-\text{var}}([0, T], V)$ such that $\log(S(X_i)) = \ell_i$. Let $\rho_M = \sum_{i=1}^N \lambda_i \delta_{\ell_i} \in \mathcal{M}(C^{1-\text{var}}([0, T], V))$. Thus, $(\tilde{\mathbf{p}}_n)_{n=0}^M$ is also orthonormal with respect to $S_* \rho_M$. \square

4. ORTHOGONAL POLYNOMIALS ON WIENER SPACE

4.1. The non-time-augmented case: non-existence. Consider a centered real-valued Gaussian random variable with variance σ^2 . Hermite polynomials are the orthogonal family associated with the measure $\mu(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}x^2}$, i.e., a sequence of polynomials $H_n(x)$, $n \geq 0$ such that

$$\int H_n(x) H_m(x) d\mu(x) = n! \delta_{n,m}.$$

They may be characterized by their generating function

$$G(t, x) = \exp\left(tx - \frac{1}{2}\sigma^2 t^2\right) = \sum_{n=0}^{\infty} \frac{1}{n!} H_n(x) t^n.$$

The first few are $H_0(x) = 1$, $H_1(x) = x$, $H_2(x) = x^2 - \sigma^2$, $H_3(x) = x^3 - 3\sigma^2 x$, and in general

$$H_n(x) = n! \sum_{k=0}^{\lfloor n/2 \rfloor} \frac{\sigma^{2k}}{k!(n-2k)!2^k} x^{n-2k}.$$

Fix a vector space V and let \mathbf{X} be a centered Gaussian vector with covariance matrix Σ , which we regard as a quadratic form on V^* . Multivariate Hermite polynomials are the block-orthogonal family associated with the pushforward measure $\mu = \mathbf{X}_* \mathbb{P} = \mathbb{P} \circ \mathbf{X}^{-1}$. They may be characterized by their generating function

$$G(t\lambda, \mathbf{x}) := \exp\left(t\lambda(\mathbf{x}) - \frac{t^2}{2}\Sigma(\lambda)\right) = \sum_{n=0}^{\infty} \frac{t^n}{n!} H_n(\lambda, \mathbf{x}) \tag{4.1}$$

Since for a given $\lambda \in V^*$ the random variable $\lambda(\mathbf{X})$ is Gaussian with variance $\Sigma(\lambda)$ it follows that the polynomials are the same univariate ones as above, evaluated at $\lambda(\mathbf{x})$ with variance $\Sigma(\lambda)$.

From the identity $\mathbb{E}[e^{\lambda(\mathbf{X})}] = e^{\frac{1}{2}\Sigma(\lambda)}$ it follows that

$$\mathbb{E}[G(s\lambda, \mathbf{X})G(t\lambda', \mathbf{X})] = e^{\frac{1}{2}stB(\lambda, \lambda')}$$

where $B(\lambda, \lambda') = \frac{1}{2}(\Sigma(\lambda + \lambda') - \Sigma(\lambda) - \Sigma(\lambda'))$ is the associated bilinear form. From this we see that

$$\int H_n(\lambda, \mathbf{x}) H_m(\lambda', \mathbf{x}) d\mu(\mathbf{x}) = n! B(\lambda, \lambda')^n \delta_{n,m}.$$

It should also be clear that $H_n(\lambda, \mathbf{x})$ depends polynomially on both λ and \mathbf{x} , homogeneously so on λ but not on x . For example, $H_0(\lambda, \mathbf{x}) = 1$, $H_1(\lambda, \mathbf{x}) = \lambda_i x^i$ and

$$H_2(\lambda, \mathbf{x}) = \lambda_i \lambda_j (x^i x^j - \Sigma^{ij}),$$

and so on, where the Einstein summation convention is in place. We note that, although these expressions depend on a choice of basis for V and the corresponding dual basis for V^* , the polynomials H_n are independent of this.

Indeed, we have intentionally omitted any reference to the dimension of \mathbf{X} since it is evident from (4.1) that the expression for H_n does not depend on it. In fact, (4.1) makes sense on any finite-dimensional space carrying a fixed quadratic form Σ and it is independent of any choice apart from this data. Multivariate Hermite polynomials are natural in a categorical sense: consider the category *Quad* of *quadratic spaces*, i.e., pairs (V, Σ_V) where V is a finite-dimensional vector space and Σ is a quadratic form. Maps between quadratic spaces are *injective isometries* (and not necessarily surjective), that is, linear maps $\varphi: U \rightarrow V$ such that $\Sigma_V(\varphi u) = \Sigma_U(u)$.

Now, let (U^*, Σ_U) and (V^*, Σ_V) be nondegenerate quadratic spaces and $\varphi: U \rightarrow V$ be such that $\varphi^*: V^* \rightarrow U^*$, $\varphi^* \lambda \mapsto \lambda \circ \varphi$ is an isometry. For fixed $\lambda \in U^*$, the map $H_n(\lambda(\cdot); \Sigma(\lambda))$ sends a monomial in U to a polynomial, so it lies in $\text{Sym}(U)^*$. By homogeneity of H_n in λ this map is then $H_n: \text{Sym}_n(U^*) \rightarrow \text{Sym}(U)^*$. The identity

$$G_V(\varphi^* \lambda, \mathbf{x}) = \exp\left(\lambda(\varphi \mathbf{x}) - \frac{1}{2}\Sigma_V(\varphi^* \lambda)\right) = G_U(\lambda, \varphi \mathbf{x})$$

implies $H_n(\varphi^* \lambda, \cdot) = H_n(\lambda, \varphi \cdot) = \varphi^* H_n(\lambda, \cdot)$ for all $n \geq 0$, that is, the diagram

$$\begin{array}{ccc} \text{Sym}(V^*) & \xrightarrow{H_V} & \text{Sym}(V)^* \\ \downarrow \text{Sym}(\varphi^*) & & \downarrow \text{Sym}(\varphi)^* \\ \text{Sym}(U^*) & \xrightarrow{H_U} & \text{Sym}(U)^* \end{array}$$

commutes. By making use of the isomorphism $V \cong V^*$ induced by the nondegenerate bilinear form B_V associated to Σ_V we obtain a commuting diagram

$$\begin{array}{ccc} \text{Sym}(V^*) & \xrightarrow{H_V} & \text{Sym}(V^*) \\ \downarrow \text{Sym}(\varphi^*) & & \downarrow \text{Sym}(\varphi^*) \\ \text{Sym}(U^*) & \xrightarrow{H_U} & \text{Sym}(U^*), \end{array}$$

This property has the advantage, already hinted above, of making any computation involving these polynomials completely independent of any choice but the covariance operator Σ .

Now, we would like a similar map for the non-time-augmented Stratonovich signature of Brownian motion. In the more restrictive setting where morphisms are taken to be bijective isometries (so that the dimension of U is fixed), this is possible, the map being given by Gram-Schmidt block orthogonalization. Unfortunately, in the more interesting case of morphisms being (not necessarily bijective) isometries this is not possible beyond degree 4 as we now show.

We turn to the d -dimensional Wiener measure, without drift for the time being. Switching notation slightly to that in Section 2, we recall the expression for the expected signature of Brownian motion [28, 41], sometimes called Fawcett’s formula, in any given basis of \mathbb{R}^d ,

$$\mathbb{E}S(W)_{0,T} = \exp\left(\frac{T}{2} \sum_{\gamma=1}^d \gamma\gamma\right) = \sum_{n=0}^{\infty} \frac{T^n}{2^n n!} \sum_{\gamma_1, \dots, \gamma_n=1}^d \gamma_1 \gamma_1 \dots \gamma_n \gamma_n. \tag{4.2}$$

For simplicity we set $T = 1$ for the remainder of this section. We would like to block-orthogonalise words in $\text{Sh}(\mathbb{R}^d)$ w.r.t. this measure. Following our considerations for Hermite, we make the following definition of *natural maps*.

Definition 4.1. We call a family of maps $H_V : T(V) \rightarrow T(V)$ *natural* if for any injective isometry $\varphi : V \rightarrow W$ the diagram

$$\begin{array}{ccc} T(V) & \xrightarrow{H_V} & T(V) \\ \downarrow T(\varphi) & & \downarrow T(\varphi) \\ T(W) & \xrightarrow{H_W} & T(W) \end{array}$$

commutes, where $T(\varphi) : T(V) \rightarrow T(W)$ is the induced map $T(\varphi)(v_1 \cdots v_n) = \varphi(v_1) \cdots \varphi(v_n)$.

Remark 4.2 (Restriction property). In particular, if $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}^D$ with $d < D$ is the inclusion map, the orthogonalization map $H_{\mathbb{R}^D}$ restricts to $H_{\mathbb{R}^d}$ on words only containing the first d coordinates.

The following is the main result of this subsection.

Theorem 4.3. For $V = \mathbb{R}^d$, define $H_V : T(V) \rightarrow T(V)$ by $H_V(w) = p_w$ for every word w , where p_w is the block-orthogonalisation endomorphism defined in (2.7) with respect to d -dimensional Brownian motion. Then H_V is not natural for $d \geq 2$.

Proof. Using the Wolfram code listed in Appendix A, we obtain for $d = 3$ and $w = 11112$, the image

$$p_w = 11112 + \frac{1}{96} 332 - \frac{1}{96} 233 - \frac{1}{96} 211 - \frac{35}{96} 112 + \frac{5}{96} 2.$$

Since p_w depends on letters not in w the map cannot be natural. For an example with $d = 2$, and the same w we obtain

$$p_w = 11112 - \frac{1}{80} 211 - \frac{29}{80} 112 + \frac{5}{96} 2$$

and we see that the coefficients have changed. □

While the above proof is short (following the computational effort in Appendix A), it perhaps does not reveal the mechanism behind the failure of naturality. We now provide an alternative, less computational, explanation for this behaviour, albeit without providing all the details.

While the map $w \mapsto p_w$ is completely determined by (2.7), computing the coefficients in this expression involves solving a linear system which might be difficult and not very insightful. In particular, it is difficult to assess the naturality of this map in its current form. Therefore, we attempt to come up with a procedure for computing it that takes the structure of the shuffle algebra and Fawcett’s formula (4.2) into account, i.e. by using the formula

$$(u, w) = \langle u \sqcup w, \mathbb{E}S(W)_{0,1} \rangle. \tag{4.3}$$

For this we develop some graphical notation. Each dot will be a placeholder for a letter in the alphabet $\{1, \dots, d\}$, and each arc will stand for a contraction, i.e. we replace the subword (α, β) formed by the contraction by a Kronecker delta $\delta_{\alpha\beta}$. Replacing indices with degree, the first three projections (which do satisfy the restriction property) can be written as

$$e_0 = \emptyset, \quad e_1 = \bullet, \quad e_2 = \bullet \bullet - \frac{1}{2} \text{arc}, \quad e_3 = \bullet \bullet \bullet - \frac{1}{4} (\text{arc} \bullet + \bullet \text{arc}). \tag{4.4}$$

Note e_n on its own does not identify an element of the shuffle algebra: this is only true once n letters are supplied. More precisely, for a word $w = \alpha_1 \dots \alpha_n$, p_w is given by e_n where we assign the letters $\alpha_1, \dots, \alpha_n$ to the n dots.

We make the following general ansatz for e_n . Since e_n must be monic, we start with the string of n dots (the word itself). We want to impose $(e_n, e_m) = 0$ for all $m < n$, and it is enough to do this for $m \equiv n \pmod{2}$. Taking the inner product as in (4.3) of the word itself with any word of degree $n - 2$, by the formula (4.2), will leave some arcs between consecutive

nodes, so it makes sense to add some multiple of such elements, where the multiple will be solved for. Taking the inner product of this element with elements of degree $n - 4$ will yield diagrams with two arcs, each of which between consecutive nodes; this is not all, however: it will also diagrams containing nested arcs, since the single arcs included in the preceding step are “skipped over”. Continuing recursively in this manner suggests the following form for e_n : begin with a string of n nodes with no pairings, and add unknown multiples of diagrams in which the pairings obey the following rules:

- The pairing is non-crossing, i.e. no two arcs intersect;
- Each node under an arc is itself paired.

This means in general, our diagrams will consist of “islands” of nodes that are paired alternated with strings of unpaired nodes. Similar diagrams have appeared in [4]. At degree 4 this becomes

$$e_4 = \dots + x_1(\text{arc}) + x_2(\text{nested arcs}) + x_3(\text{two arcs}) + y_1(\text{two arcs}) + y_2(\text{nested arcs}).$$

Solving for $(e_4, e_2) = 0 = (e_4, e_0)$ we obtain a unique solution!

$$e_4 = \dots - \frac{1}{6}(\text{arc}) - \frac{1}{6}(\text{nested arcs}) - \frac{1}{6}(\text{two arcs}) + \frac{1}{24}(\text{two arcs}) + \frac{1}{12}(\text{nested arcs}). \tag{4.5}$$

Now we ask whether this ansatz works in general. At level 5 it reads

$$e_5 = \dots + x_1(\text{arc}) + x_2(\text{nested arcs}) + x_3(\text{two arcs}) + x_4(\text{two arcs}) + y_1(\text{two arcs}) + y_2(\text{nested arcs}) + y_3(\text{two arcs}) + y_4(\text{nested arcs}) + y_5(\text{two arcs}). \tag{4.6}$$

Now we claim that the linear system arising from the constraints $(e_5, e_3) = 0 = (e_5, e_1)$ does not have a solution. Indeed, the constraints are obviously equivalent to $(e_5, w_3) = 0 = (e_5, w_1)$, where w_n denotes a word of length n .

On the one hand, the equation $(e_5, w_1) = 0$ yields

$$0 = \dots + \left(y_5 + \frac{1}{4}x_3\right)(\text{diagram with red dot}) + \dots \tag{4.7}$$

where \dots represents terms with different full pairings and the red dot indicates the letter not in the original word. Taking an alphabet with at least three letters and choosing the dot assignments as $\alpha\beta\gamma\gamma\beta\alpha$ for the six dots in (4.7), we see that all terms in \dots vanish and the constraints imply $y_5 + \frac{1}{4}x_3 = 0$.

On the other hand, the equation $(e_5, w_3) = 0$ gives

$$0 = \left(\frac{1}{48} + \frac{1}{6}x_3\right)(\text{diagram with red dots}) + \dots + \left(\frac{1}{4}y_5 + \frac{1}{12}x_3\right)(\text{diagram with red dots}) + \dots$$

It can be checked that these are the only terms producing the corresponding diagrams. The difference in the coefficients comes from the fact that, in each cases, there are different numbers of shuffles in (4.3) that yield the diagram, and also the number of contractions is different. This yields $y_5 + \frac{1}{3}x_3 = 0$ and $\frac{1}{8} + x_3 = 0$ which contradicts the earlier equation $y_5 + \frac{1}{4}x_3 = 0$.

This proves that the general ansatz as in (4.6) is not possible. What is missing to make the above argument an alternative proof of Theorem 4.3 is a justification for why naturality would imply that p_w must be given by e_n as in the above ansatz.

Every natural transformation in the sense of Definition 4.1 must be $O(V)$ -equivariant. The later class of maps is indexed by Brauer diagrams [13] (see also the more modern article [38]). Roughly speaking, there are four types of basic diagrams (maps) [38, Theorem 2.6] yielding all possible transformations by composition and tensorization. These are the identity map, transposition of two tensor factors; *caps*, corresponding to contractions $\alpha\beta \mapsto \delta_{\alpha\beta}$ as above and *cups*, corresponding to maps of the form

$$\mathbb{R} \ni \lambda \mapsto \lambda \sum_{\alpha=1}^d \alpha\alpha,$$

where d denotes the dimension of the underlying vector space. Note that of these four, only cups depend on the dimension and are therefore not natural, hence our ansatzes in (4.4)-(4.6) include only identities and caps. The choice of including only noncrossing diagrams is motivated by Fawcett’s formula, but we do not give a non-computational proof that crossing diagrams do not appear.

We have, however, analysed computationally the same ansatz as (4.4)-(4.6) where we now include all possible pairings up to level 5. This extended ansatz yields the same solutions up to degree four: the extra degrees of freedom are set to zero, and at degree 5 the system still contains a subset of inconsistent equations. See Appendix A for Wolfram code listings exploring this.

4.2. The time-augmented case: Itô orthogonal polynomials. We have seen that, without including time, the block orthogonalisation with respect to the signature of Brownian motion is not natural in the sense of Theorem 4.3. We now show that including time as a coordinate solves this problem, as it enables the use of the Itô integral, whose product rule involves quadratic variation corrections. Let V be an Euclidean vector space, for which we temporarily fix an orthonormal frame, i.e. an isomorphism $V \cong \mathbb{R}^d$; we will return to coordinate-free considerations later on, and recall that even without the choice of a basis the inner product yields a canonical identification $V \cong V^*$. Write $\tilde{V} := \mathbb{R} \oplus V$. Elements of $T(\tilde{V})$ can be identified with linear combinations of words in the alphabet $[d]_0 := \{0, 1, \dots, d\}$; write $[d]_0^\bullet$ for the set of such words. Note the distinction between the letter 0 and the empty word \emptyset , which spans $\mathbb{R} = \tilde{V}^{\otimes 0}$. Recall the *quasi-shuffle* product on $T(\tilde{V})$, given recursively on words as follows and extended bilinearly:

$$u\alpha \widehat{\sqcup} v\beta = (u \widehat{\sqcup} v\beta)\alpha + (u\alpha \widehat{\sqcup} v)\beta + (u \widehat{\sqcup} v)[\alpha, \beta] \quad \text{with} \quad [\alpha, \beta] = \mathbb{1}_{\alpha \neq \emptyset} 0. \quad (4.8)$$

Denote $\widehat{\text{Sh}}(\tilde{V}) = T(\tilde{V})$ endowed with this product. We consider two gradings on this space: $\#w$ denotes the number of letters in the word w , and

$$|w| := \#w + \text{number of } 0\text{'s in } w. \quad (4.9)$$

Given a V -valued Wiener process B , write $\tilde{B}_t := (t, B_t)$ for its time augmentation in the 0th coordinate. We denote $\widehat{S}(\tilde{B})$ its Itô signature, i.e. (2.2) but where the integrals are defined by Riemann and Itô integration. Recall (e.g. [6]) the product rule (2.3) is replaced by

$$\langle u, \widehat{S}(\tilde{B}) \rangle \langle v, \widehat{S}(\tilde{B}) \rangle = \langle u \widehat{\sqcup} v, \widehat{S}(\tilde{B}) \rangle. \quad (4.10)$$

The bracket in (4.8) then indexes quadratic variation, i.e. if $v_1, v_2 \in V$

$$\langle [v_1, v_2], dB_t \rangle = \varrho_{v_1, v_2} dt \quad (4.11)$$

where $\varrho_{v_1, v_2} = \mathbb{E}[B_1^{v_1} B_1^{v_2}]$ denotes the (constant) correlation of the components of B in the directions v_1, v_2 . Following the same logic of (2.5), we define the inner product on $\widehat{\text{Sh}}(\tilde{V})$

$$(u, v)_{\widehat{\sqcup}} := \mathbb{E} \langle u, \widehat{S}(\tilde{B})_{0,T} \rangle \langle v, \widehat{S}(\tilde{B})_{0,T} \rangle = \langle u \widehat{\sqcup} v, \mathbb{E} \widehat{S}(\tilde{B})_{0,T} \rangle. \quad (4.12)$$

The introduction of time as a coordinate has the consequence of making the inner product degenerate, for example $\langle 0 - T\emptyset, \widehat{S}(\tilde{B})_{0,T} \rangle = 0$ a.s. The next proposition identifies a complement to the nullspace (cf. [27, Theorem 3.9] for a linear independence statement).

Proposition 4.4. *Let N be the nullspace of $(\cdot, \cdot)_{\widehat{\sqcup}}$ and denote $\widehat{\text{Sh}}^\circ(\tilde{V})$ the vector subspace of $\widehat{\text{Sh}}(\tilde{V})$ generated by words that do not end in a 0. Then $\widehat{\text{Sh}}(\tilde{V}) = N \oplus \widehat{\text{Sh}}^\circ(\tilde{V})$.*

Proof. Since $(\cdot, \cdot)_{\widehat{\sqcup}}$ is just the L^2 inner product applied to the evaluation of a word on $\widehat{S}(\tilde{B})_{0,T}$, it suffices to show that (i) if $\rho \in \widehat{\text{Sh}}^\circ(\tilde{V})$ such that $\mathbb{E} \langle \rho, \widehat{S}(\tilde{B})_{0,T} \rangle^2 = 0$ then $\rho = 0$, and (ii) that for any $w \in [d]_0^\bullet$ there exists a $\rho \in \widehat{\text{Sh}}^\circ(\tilde{V})$ such that $\langle w0, \widehat{S}(\tilde{B})_{0,T} \rangle = \langle \rho, \widehat{S}(\tilde{B})_{0,T} \rangle$.

The claim (ii) follows inductively on the number of trailing 0's, since, by the (quasi-)shuffle relations, $\langle w0, \widehat{S}(\tilde{B})_{0,T} \rangle$ can be expressed as the product $\langle 0, \widehat{S}(\tilde{B})_{0,T} \rangle \langle w, \widehat{S}(\tilde{B})_{0,T} \rangle$ minus terms with fewer trailing zeros, and $\langle 0, \widehat{S}(\tilde{B})_{0,T} \rangle = T \in \mathbb{R}$. In order to prove (i), we first show that if the function

$$[0, T] \ni t \mapsto \langle \sigma, \widehat{S}(\tilde{B})_{0,t} \rangle$$

vanishes identically on some $\sigma \in \widehat{\text{Sh}}(\tilde{V})$, then $\sigma = 0$. Proceed by induction on the maximum length n of a word in the linear combination for σ . If $n = 0$, $\langle \sigma, \widehat{S}(\tilde{B})_{0,t} \rangle \equiv \sigma \in \mathbb{R}$. For $n > 0$, write $\sigma = \lambda\emptyset + \sum_i \lambda_{0,i} w_{0,i} 0 + \sum_{\alpha=1}^d \sum_j \lambda_{\alpha,i\alpha} w_{\alpha,i\alpha}$. The process $\langle \sigma, \widehat{S}(\tilde{B})_{0,t} \rangle \equiv 0$ is then equal to a constant term plus a linear combination of Riemann and Itô integrals, all of whose integrands must vanish identically, by Doob-Meyer, independence of components, and the fact that if $\int_0^t H_s d\mathbf{B}_s^\alpha$ or $\int_0^t H_s ds$ vanishes for all $t \in [0, T]$ then H must also vanish identically. Therefore $\lambda = 0$, $\sum_j \lambda_{0,i} w_{0,i} 0 = 0$ and $\sum_j \lambda_{\alpha,i\alpha} w_{\alpha,i\alpha} = 0$ for all α and the induction is complete. Now, assume $\rho = \mu\emptyset + \sum_{\alpha=1}^d \sum_j \lambda_{\alpha,j} w_j \alpha$ is as in the statement of (i) above. By the Itô isometry

$$0 = \mathbb{E} \langle \rho, \widehat{S}(\tilde{B})_{0,T} \rangle^2 = \sum_{\alpha=1}^d \int_0^T \mathbb{E} \langle \sum_j \lambda_{\alpha,j} w_j, \widehat{S}(\tilde{B})_{0,t} \rangle^2 dt$$

and the integrand does not vanish identically, since the process $t \mapsto \langle \sum_j \lambda_{\alpha,j} w_j, \widehat{S}(\tilde{B})_{0,t} \rangle$ does not, as just proved. \square

Write $[d]_0^\circ \subset [d]_0^\bullet$ for the subset of words that do not end in a 0, so $\text{span}[d]_0^\circ = \widehat{\text{Sh}}^\circ(\tilde{V})$. Using Itô integration has the benefit that, by independence and orthogonality of Wiener chaos, many pairs of words are already orthogonal. For $u, v \in [d]_0^\circ$

write $u \sim_0 v$ if u and v are equal after stripping away zeros and leaving other letters in their place. Observe that (by the martingale property of Itô integrals)

$$\mathbb{E}\widehat{S}(\widetilde{B})_{0,T} = \sum_{n=0}^{\infty} \frac{T^n}{n!} 0^n \tag{4.13}$$

and therefore $(u, v)_{\widehat{\square}} = 0$ whenever $u \not\sim_0 v$, since in this case $u \widehat{\square} v$ will be a linear combination of words none of which are of the form 0^n . Moreover, if $u \sim_0 v$, $(u, v)_{\widehat{\square}}$ will only depend on the number and position of the non-zero letters, i.e. it will equal the inner product of the words in the binary alphabet $\{0, 1\}$ in which all non-zero entries in u and v are replaced with a 1, for which we compute

$$(0^{i_1} 1 0^{j_2} \dots 0^{i_k} 1, 0^{i_1} 1 0^{j_2} \dots 0^{i_k} 1)_{\widehat{\square}} = \frac{T^{i+j+k}}{(i+j+k)!} \prod_{r=0}^k \binom{i_r + j_r}{i_r}, \quad \sum_{l=1}^k i_l = i, \quad \sum_{l=1}^k j_l = j. \tag{4.14}$$

in which i_l, j_l may be 0. Define a linear order $<_0$ on each equivalence class mod \sim_0 as follows:

- 1 If $u \sim_0 v$ and $\#u < \#v$ ($\Leftrightarrow |u| < |v| \Leftrightarrow u$ has fewer zeros than v), then $u <_0 v$;
- 2 If $u \sim_0 v$ and $\#u = \#v$, then $u <_0 v$ if u is less than v in the lexicographic order.

The following definition is similar to (2.7) but adapted to the current setting with 0 playing a special role.

Definition 4.5 (Itô orthogonal basis). For $u \in [d]_0^\circ$ define $\widehat{\rho}_u$ by performing Gram-Schmidt orthogonalisation along its equivalence class mod \sim_0 , according to its linear order:

$$\widehat{\rho}_w = w - \sum_{\substack{v \sim_0 w \\ v <_0 w}} \frac{(w, \widehat{\rho}_v)_{\widehat{\square}}}{(\widehat{\rho}_v, \widehat{\rho}_v)_{\widehat{\square}}} \widehat{\rho}_v.$$

Each $\widehat{\rho}_w$ is a linear combination of other words that contain the same non-zero letters (as a subword, in fact), with the same number of zeros or fewer. We compute the terms of $\widehat{\rho}$ up to words of length 3. We use letters α, β, \dots to denote letters in $[d]$.

$$\begin{aligned} \widehat{\rho}_\emptyset &= \emptyset, & \widehat{\rho}_0 &= 0, & \widehat{\rho}_{01} &= 01 - \frac{1}{2}1, & \widehat{\rho}_{001} &= 001 - \frac{1}{2}01 + \frac{1}{12}1 \\ \widehat{\rho}_{11} &= 11, & \widehat{\rho}_{011} &= 011 - \frac{1}{3}11, & \widehat{\rho}_{101} &= 101 + \frac{1}{2}011 - \frac{1}{2}11, & \widehat{\rho}_{111} &= 111. \end{aligned}$$

While we have described the orthogonal basis $\{\widehat{\rho}_w \mid w \in [d]_0^\circ\}$ in terms of a basis on V , it comes from an intrinsic orthogonalisation map. Indeed, we have decomposed

$$\widehat{\text{Sh}}^\circ(\widetilde{V}) = \bigoplus_{k=0}^{\infty} W^k, \quad W^k = \bigoplus_{i=0}^{\infty} \underbrace{\bigoplus_{i_1+\dots+i_k=i} W_{i_1 \dots i_k}^k}_{|\cdot| = k+2i, \# = k+i} \tag{4.15}$$

with W^k the space generated by all words with k non-zero letters (ending in a non-zero letter) and $W_{i_1 \dots i_k}^k$ its subspace of generated by words of the form $0^{i_1} v_1 0^{i_2} \dots 0^{i_k} v_k$ with $v_l \in V$ and the multiplicities i_l possibly zero. The first direct sum is orthogonal, since W^k maps to the k^{th} Wiener chaos, but the other two are not. For each fixed k , the blocks $W_{i_1 \dots i_k}^k$ are linearly ordered according to the order on sequences of integers given by $i_1 \dots i_k < i'_1 \dots i'_k$ if $\sum_l i_l < \sum_l i'_l$ or if $\sum_l i_l = \sum_l i'_l$ and for some $h \leq k$, $i_l = i'_l$ for $l < h$ and $i_h < i'_h$. We then perform block-orthogonalisation $\widehat{\rho}$ according to this order on the blocks $W_{i_1 \dots i_k}^k$: $\widehat{\rho}_w$ is just the orthogonal projection of w onto the direct complement of all preceding blocks in W^k . Denote the direct sum of these maps $\widehat{\rho}^V: \widehat{\text{Sh}}^\circ(\widetilde{V}) \rightarrow \widehat{\text{Sh}}^\circ(\widetilde{V})$. A basis of V additionally yields an orthogonal basis of each block $W_{i_1 \dots i_k}^k$, and orthogonality is preserved by $\widehat{\rho}$, yielding an orthogonal basis. Notice that only a basis of V , not a frame as in the more general setting (2.7), is necessary.

Theorem 4.6 (Natural orthogonalisation of Wiener functionals). With \mathcal{F} the sigma-algebra generated by $(B_t)_{t \in [0, T]}$, $Y \in L^2(\mathcal{F})$ can be represented as the series with pairwise orthogonal summands, convergent in L^2

$$Y = \sum_{w \in [d]_0^\circ} \frac{\mathbb{E}[Y(\widehat{\rho}_w, \widehat{S}(\widetilde{B})_{0,T})]}{(\widehat{\rho}_w, \widehat{\rho}_w)_{\widehat{\square}}} (\widehat{\rho}_w, \widehat{S}(\widetilde{B})_{0,T}).$$

Moreover, \widehat{p} is a natural orthogonalisation according to the functor $V \mapsto \widehat{\text{Sh}}^\circ(\widetilde{V})$, i.e. for any injective isometry of Euclidean spaces $\varphi: U \rightarrow V$, the diagram

$$\begin{array}{ccc} \widehat{\text{Sh}}^\circ(\widetilde{U}) & \xrightarrow{\widehat{p}^U} & \widehat{\text{Sh}}^\circ(\widetilde{U}) \\ \downarrow \widehat{\text{Sh}}^\circ(\varphi) & & \downarrow \widehat{\text{Sh}}^\circ(\varphi) \\ \widehat{\text{Sh}}^\circ(\widetilde{V}) & \xrightarrow{\widehat{p}^V} & \widehat{\text{Sh}}^\circ(\widetilde{V}) \end{array}$$

(with $\widehat{\text{Sh}}^\circ(\varphi) := T(\mathbb{1} \oplus \varphi)|_{\widehat{\text{Sh}}^\circ(\widetilde{U})}$) commutes.

Proof. The first part follows directly from Corollary 2.3 (or rather the same proof; since Theorem 2.4 or [14] apply to time-augmented Brownian motion), (4.4) and the preceding discussion. Let $\Phi := \widehat{\text{Sh}}^\circ(\varphi)$. It follows from the preceding discussion that the decomposition into blocks $W_{i_1 \dots i_k}^k$ is preserved by Φ , and moreover Φ acts on $W_{i_1 \dots i_k}^k = W_{i_1 \dots i_k}^k(U) \cong U^{\otimes k}$ by $\varphi^{\otimes k}$. Since \widehat{p} operates on each block W^k independently, we have reduced the problem to proving, for each $i, k \geq 0$, commutation of the squares

$$\begin{array}{ccc} \bigoplus_{i_1 + \dots + i_k = i} U^{\otimes k} & \xrightarrow{\widehat{p}^U} & \bigoplus_{i_1 + \dots + i_k = i} U^{\otimes k} \\ \downarrow \varphi^{\otimes k} & & \downarrow \varphi^{\otimes k} \\ \bigoplus_{i_1 + \dots + i_k = i} V^{\otimes k} & \xrightarrow{\widehat{p}^V} & \bigoplus_{i_1 + \dots + i_k = i} V^{\otimes k}. \end{array}$$

This follows from the fact that φ preserves the bracket in (4.8), i.e. $[\varphi(u), \varphi(u')] = [\varphi(u), \varphi(u')]$ and therefore preserves inner products in W^k :

$$(0^{i_1} \varphi(u_1) 0^{i_2} \dots 0^{i_k} \varphi(u_k), 0^{j_1} \varphi(u'_1) 0^{j_2} \dots 0^{j_k} \varphi(u'_k)) = (0^{i_1} u_1 0^{i_2} \dots 0^{i_k} u_k, 0^{j_1} u'_1 0^{j_2} \dots 0^{j_k} u'_k).$$

The whole Gram-Schmidt process Definition 4.5 is thus preserved and the claim follows. □

Remark 4.7 (Stratonovich orthogonalisation). A similar orthogonalisation map can be obtained when working with the Stratonovich signatures. We only sketch this briefly, since it is easier to interpret terms in the Itô setting. Recall that the Hu-Meyer formulae [35] (later generalised to the Hoffman’s exponential [34] in a different context) establish natural algebra isomorphisms

$$\exp^V: \text{Sh}(\widetilde{V}) \rightarrow \widehat{\text{Sh}}(\widetilde{V}), \quad \log^V: \widehat{\text{Sh}}(\widetilde{V}) \rightarrow \text{Sh}(\widetilde{V})$$

inverses of each other. A quick inspection of these maps reveals that they respect the direct sum of Proposition 4.4, since the bracket does not act on the 0 coordinate. We use $\langle \cdot, \cdot \rangle_{\sqcup}$ to denote the inner product on $\text{Sh}(\widetilde{V})$ in which evaluation is against the Stratonovich signature $S(\widetilde{W})_{0,T}$. Then

$$\langle u, v \rangle_{\sqcup} = \langle u \sqcup v, S(\widetilde{B})_{0,T} \rangle = \langle \exp^V(u) \sqcup \exp^V(v), \widehat{S}(\widetilde{B})_{0,T} \rangle = \langle \exp^V(u), \exp^V(v) \rangle_{\widehat{\sqcup}},$$

i.e. \exp^V (and \log^V) define isometries. Therefore, setting

$$q^V: \text{Sh}(\widetilde{V}) \rightarrow \text{Sh}(\widetilde{V}), \quad q^V = \exp^V \circ \widehat{p}^V \circ \log^V$$

all properties of \widehat{p} carry over to the Stratonovich-shuffle case.

Remark 4.8 (Hermite and Legendre). We observe that \widehat{p} embeds both Hermite and Legendre polynomials: the former correspond to $n\widehat{p}_{1^n}$, the latter (shifted to the interval $[0, T]$) to $n\widehat{p}_{0^n}$.

Remark 4.9 (Comparison with Wiener chaos). Since orthogonalisation map \widehat{p} works largely because of orthogonality of Wiener chaos, it is interesting to probe the link with this decomposition further. Recall that, letting $\mathcal{H} = L^2([0, T], \mathbb{R}^d) = L^2([0, T] \times [d])$ there exists an isometry given by multiple Wiener integration (see e.g. [46])

$$\bigoplus_{n=0}^{\infty} L^2_{\text{sym}}(\underbrace{([0, T] \times [d]) \times \dots \times ([0, T] \times [d])}_n) = \bigoplus_{n=0}^{\infty} \mathcal{H}^{\odot n} \cong L^2(\mathcal{F}) \tag{4.16}$$

where the subscript sym refers to symmetry of the functions in the n factors. A straightforward way of obtaining an orthogonal basis of $L^2(\mathcal{F})$ is that of fixing one of $L^2[0, T]$ (Fourier, Legendre, etc.), lifting it to one $\{\varphi_m\}_{m \in \mathbb{N}}$ of $L^2([0, T], \mathbb{R}^d)$, and of taking symmetric powers of it. This yields a representation

$$Y = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \sum_{m_1 + \dots + m_n = m} \lambda_{m_1, \dots, m_n} \int_{[0, T]^n} \varphi_{m_1, \alpha_1}(t_1) \dots \varphi_{m_n, \alpha_n}(t_n) dW_{t_1}^{\alpha_1} \dots dW_{t_n}^{\alpha_n}. \tag{4.17}$$

with implicit summation on each α_l and coefficients λ symmetric in the lower indices. This basis however depends on two parameters n and m , and in order to obtain a one-parameter basis, one must make a choice of how to trade off truncation in n with truncation in m . Notice, moreover, that quantities of interest in differential equations, such as the Lévy areas $\frac{1}{2} \int_0^T (B^\alpha dB^\beta - B^\beta dB^\alpha)$, which are not the Wiener integral of symmetric kernels in the sense above, have infinite

expansion w.r.t. (4.17). Definition 4.5 is adapted to Wiener chaos but sidesteps these issues, thanks to the time-ordered nature of the signature.

Pushing similar considerations a little further we are able to obtain a quick proof of Theorem 2.4 in the case of time-augmented Brownian motion and $p = 2$. Recall that, even though $\mathcal{F} = \mathcal{G}$ by Remark 2.1, the proposition does not show that linear functions on $S(B)_{0,T}$ (without time-augmentation) is dense in $L^2(\mathcal{F})$, and Theorem 2.4 is needed for this.

Proposition 4.10 (L^2 -density of signature functionals). *Linear functions on $S(\widetilde{B})_{0,T}$ are dense in $L^2(\mathcal{F})$.*

Proof. By the Itô-Stratonovich conversion formulae for iterated integrals of Brownian motion with time [35], we may equivalently prove the statement for linear functions on $\widehat{S}(\widetilde{B})_{0,T}$. Since polynomials are dense in $L^2[0, T]$, by (4.16) it suffices to show that for any $n, m_1, \dots, m_n \in \mathbb{N}$ and any $\alpha_1, \dots, \alpha_n \in [d]$ the multiple Wiener integral

$$\int_{[0, T]^n} t_1^{m_1} \dots t_n^{m_n} dW_{t_1}^{\alpha_1} \dots dW_{t_n}^{\alpha_n}$$

can be expressed as a linear function of $\widehat{S}(\widetilde{B})_{0,T}$. We express it as, up to a factor of $m_1! \dots m_n!$

$$\begin{aligned} & \int_{[0, T]^n} \left(\int_{\Delta^{m_1}[0, t_1]} ds_{1,1} \dots ds_{1,m_1} \right) \dots \left(\int_{\Delta^{m_n}[0, t_n]} ds_{n,1} \dots ds_{n,m_n} \right) dW_{t_1}^{\alpha_1} \dots dW_{t_n}^{\alpha_n} \\ &= \int_{[0, T]^n \times (\Delta^{m_1}[0, t_1] \times \dots \times \Delta^{m_n}[0, t_n])} ds_{1,1} \dots ds_{1,m_1} \dots ds_{n,1} \dots ds_{n,m_n} dW_{t_1}^{\alpha_1} \dots dW_{t_n}^{\alpha_n} \end{aligned}$$

where for a family of sets $(B_a)_{a \in A}$ we denote $A \times B_a := \{(a, b) \mid a \in A, b \in B_a\}$. Writing $[0, T]^n = \bigsqcup_{\sigma \in \mathfrak{S}_n} \sigma_* \Delta^n[0, T]$, we reduce the problem of expressing (up to reordering) $\Delta^n[0, T] \times (\Delta^{m_1}[0, t_1] \times \dots \times \Delta^{m_n}[0, t_n])$ as a disjoint union of simplices over $[0, T]$. By treating the two-factor case

$$\begin{aligned} & \Delta^i[0, a] \times \Delta^j[0, b] \\ &= \bigsqcup_{k=1}^j \{0 < u_1 < \dots < u_i < a\} \times \{0 < u_{i+1} < \dots < u_{i+k} < a < u_{i+k+1} < \dots < u_{i+j} < b\} \\ &= \bigsqcup_{k=1}^j \bigsqcup_{\sigma \in \text{Sh}(i, k)} \{0 < u_{\sigma(1)} < \dots < u_{\sigma(i+k)} < a < u_{\sigma(i+k+1)} < \dots < u_{\sigma(i+j)} < b\} \end{aligned}$$

we reduce, by induction, to sets of the form

$$\Delta^n[0, T] \times (\Delta^i[0, t_1] \times \Delta^k[t_1, t_2] \times \dots \times \Delta^h[t_{n-1}, t_n]) = \Delta^{n+(i+\dots+i_n)}[0, T]$$

and the proof is complete. \square

4.3. Numerical experiments. In this subsection we perform numerical experiments on the orthogonalised Itô signature developed in the previous. The first necessary step is to obtain the Gram-Schmidt-orthogonalised Itô-signature features, up to a given inhomogeneous degree (that is with the drift coordinate counting double). We use the package `Signax` [58] for computing signatures and [49] for algebraic manipulation on the tensor algebra. Since `Signax` (as well as all other packages that compute signatures), when called on Brownian motion, compute the Stratonovich signature, we must first convert to Itô form. This is done symbolically by implementing the Hoffman logarithm (see Remark 4.7). Note that, on finitely discretised time series, this is not equivalent to computing Itô integrals as left-endpoint Riemann sums, but they are equivalent in the limit of vanishing mesh size. In order to perform the Gram-Schmidt orthogonalisation, the inner product $(u, v)_{\widehat{\square}}$ (4.12) is computed on binary words as in (4.14). The Gram-Schmidt basis (4.5) is then calculated, still in the binary case, and only then “mapped on” to the case of general d as described in Section 4.2. This is an expression of the naturality of the orthogonal basis Theorem 4.6 and avoids the explicit computation of the (quasi-)shuffle product or Monte Carlo evaluation which would normally be necessary for (2.3), significantly speeding up computation of the inner product (4.10). For further details, we refer to our implementation [17] and its documentation. We are able to check that the orthogonalisation achieves the required goal, see Figure 2.

We consider two related but distinct tasks. In *functional expansion*, the function on paths is known and the task is to approximate it as a signature expansion. This can often be done independently of data; for example if the function is an SDE (1.3), one can use a numerical method like (1.4) iterated over many intervals. In *functional regression*, the function on paths is not known or hard to expand analytically, rather we have access to i.i.d. input-output pairs, but the goal is similar. We will use orthogonal polynomials on Wiener space in a similar way for both, but the methodologies we compare with for each are distinct.

We consider a linear SDE

$$dY^k = A_{\alpha i}^k Y^i dW^\alpha, \quad Y_0 = y_0, \tag{4.18}$$

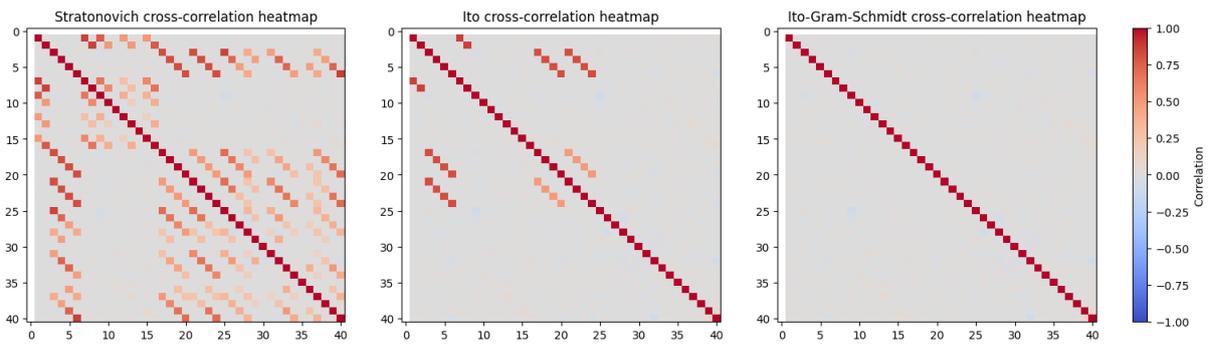


FIGURE 2. Comparison of correlation heatmaps for flattened signatures, on the subspace making $(\cdot, \cdot)_{\hat{\mathbb{H}}}$ is non-degenerate (4.4), computed over 100k 2-dimensional time-augmented sample Brownian paths, with $T = 1$ and 1k grid points. The Stratonovich signature features are far from orthogonal, the Itô ones are much sparser but still not orthogonal (because of residual correlations inside each Wiener chaos), and finally their Gram-Schmidt orthogonalisation is verified to be fully orthogonal (modulo numerical errors). This and similar checks can be performed with the notebook [17, orth_checks.ipynb].

which is one of the rare cases for which the stochastic Taylor expansion (1.4) converges on a single interval [43, §4.2]. We compare the performance of the Taylor scheme with that of evaluating the truncated series Theorem 4.6 on a sample of Brownian paths, see Figure 3. We observe convergence of both methods, with the Taylor method lagging behind the L^2 for low degrees, and catching up to the orthogonal expansion at higher degrees, especially for lower sample sizes (which only affects the orthogonal expansion). For a non-linear SDE whose solution lies in L^2 , Theorem 4.6 still applies but examples such as (1.1) show that one cannot expect convergence of the Taylor scheme on one interval, and thus iterating the method would be necessary; this however is not a signature expansion as usually intended in the machine learning literature.

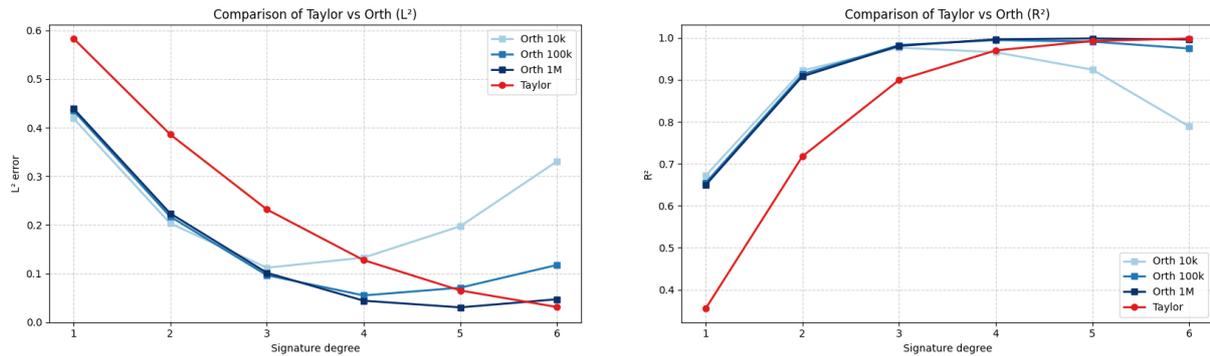


FIGURE 3. Comparison of Taylor and Orth models: out-of-sample L^2 error (left) and coefficient of determination (R^2 , right) across different dataset sizes. The Brownian motion is taken to have dimension $d = 2$, Y is scalar, and errors/ R^2 are averaged over 10 random choices of the matrix A normalised to have Euclidean norm 1.

In the next example we consider the a scalar Black Scholes model, and learn the two functions: an at-the-money call option and a lookback option, i.e. just the maximum of the path (a difficult case for functional expansions [26, Example 2]), see Figure 4. Note that unlike in the literature on applications of signatures to finance (e.g. [48]), we are learning/expanding the payoff in terms of the underlying Brownian motion, not the price path. For the latter, we would need to orthogonalise the signature of paths drawn from geometric Brownian motion. We observe convergence of both methods, both in the sample size and truncation level, although the Monte Carlo estimator for the orthogonal expansion, i.e. the truncated series in Theorem 4.6, appears to converge more slowly than the OLS estimator (2.10) in terms of the sample size, especially at higher degree. Notice that these are two distinct estimators for the same quantity, $\Pi_N Y$. In particular, for the orthogonal expansion estimator the difference between in- and out-of sample error is less marked, since it is not the solution to a data-dependent optimisation problem.

Evaluating the expansion Theorem 4.6 is linear in both sample size M and number of features $D = \frac{d^{N+1}-1}{d-1}$, while performing the matrix inversion in (2.10) has complexity which is cubic in either K or M depending on whether it is evaluated in the primal or dual formulation. In practice, the bottleneck lies in the overhead costs of computing the signatures of large

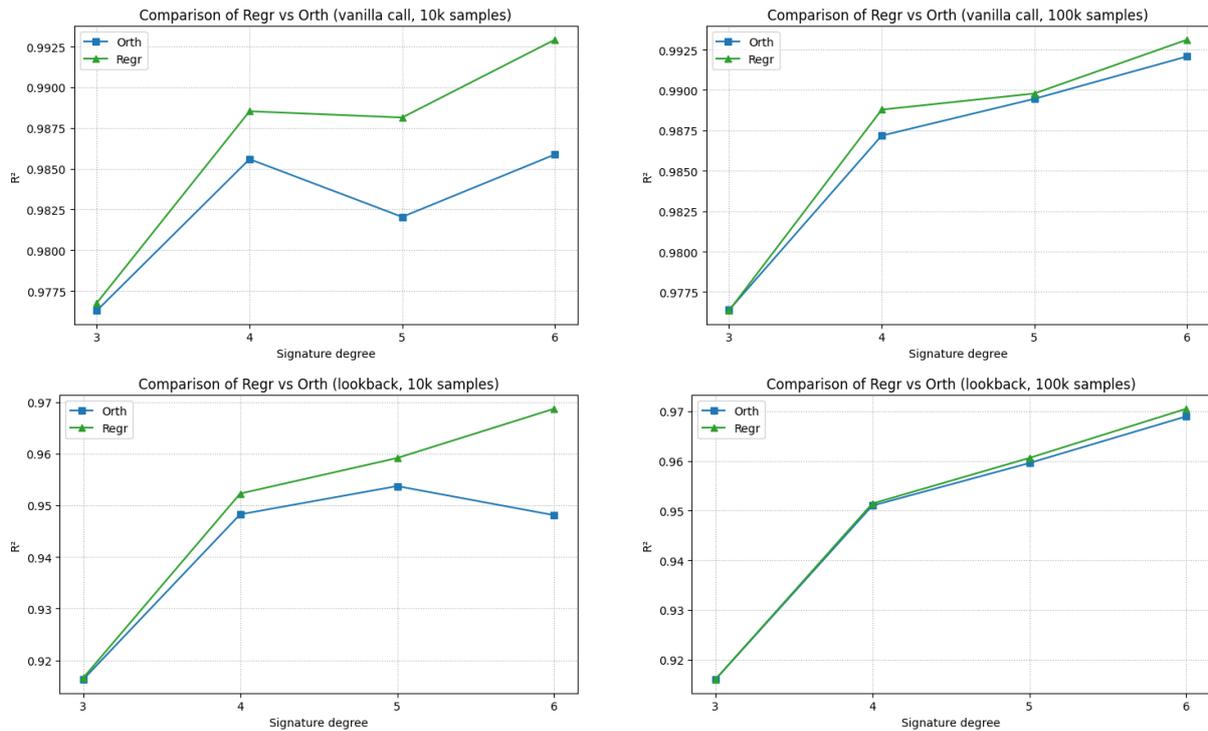


FIGURE 4. Comparison of coefficients of determination for OLS regression (Regr) on the truncated signature (with non-orthogonal coordinates, out-of-sample) and orthogonal signature expansion (Orth). The scalar Black Scholes model has parameters $\sigma = 0.2$, $\mu = 0$, $S_0 = 1$, and the call option is struck at $K = 1$ at time $T = 1$. We observed worse performance for both models for OTM options, and for ITM options Orth was performing worse than Regr.

samples of paths. Also, when sampling from a fixed measure, the data matrix could be computed once and for all and re-used across many different problems. Even when performing regression, however, using the orthogonal basis rather than the ordinary basis has advantages. Orthogonality means coefficients can be learnt incrementally in degree, unlike for non-orthogonal features. It can also improve conditioning, see [57] for a comparison in the case of polynomials (ill conditioning of Vandermonde matrices is a well known phenomenon [47]). Finally, expanding a whole collection of payoffs as in Theorem 4.6 has the benefit that covariances can be more efficiently estimated by truncating the series

$$\begin{aligned} \mathbb{E}[Y_1 Y_2] &= \sum_{w \in [d]_0^0} \frac{\mathbb{E}[Y_1 \langle \hat{\rho}_w, \hat{S}(\tilde{B})_{0,T} \rangle] \mathbb{E}[Y_2 \langle \hat{\rho}_w, \hat{S}(\tilde{B})_{0,T} \rangle]}{(\hat{\rho}_w, \hat{\rho}_w)_{\hat{\Pi}}} \langle \hat{\rho}_w, \hat{S}(\tilde{B})_{0,T} \rangle^2 \\ &= \sum_{w \in [d]_0^0} \frac{\mathbb{E}[Y_1 \langle \hat{\rho}_w, \hat{S}(\tilde{B})_{0,T} \rangle] \mathbb{E}[Y_2 \langle \hat{\rho}_w, \hat{S}(\tilde{B})_{0,T} \rangle]}{(\hat{\rho}_w, \hat{\rho}_w)_{\hat{\Pi}}} \end{aligned}$$

This could have applications, for example, in stochastic portfolio theory [22].

REFERENCES

- [1] Damien Ackerer and Damir Filipović. Option pricing with orthogonal polynomial expansions. *Math. Finance*, 30(1):47–84, 2020.
- [2] Daniel Alpay, Palle E. T. Jorgensen, and David P. Kimsey. Moment problems in an infinite number of variables. *Infinite Dimensional Analysis, Quantum Probability and Related Topics*, 18(04):1550024, December 2015.
- [3] George E. Andrews and Richard Askey. Classical orthogonal polynomials. *Polynômes orthogonaux et applications*, Proc. Laguerre Symp., Bar-le-Duc/France 1984, Lect. Notes Math. 1171, 36-62 (1985)., 1985.
- [4] Michael Anshelevich. Appell polynomials and their relatives. *Int. Math. Res. Not.*, 2004(65):3469–3531, 2004.
- [5] F. Bach. *Learning Theory from First Principles*. Adaptive Computation and Machine Learning series. MIT Press, 2024.
- [6] Fabrice Baudoin. *An introduction to the geometry of stochastic flows*. Imperial College Press, London, 2004.
- [7] Christian Bayer, Luca Pelizzari, and John Schoenmakers. Primal and dual optimal stopping with signatures. *arXiv e-prints: arXiv:2312.03444*, December 2023.
- [8] Christian Bayer and Josef Teichmann. The Proof of Tchakaloff’s Theorem. *Proceedings of the American Mathematical Society*, 134(10):3035–3040, 2006.
- [9] Christian Berg and J. P. Reus Christensen. Density questions in the classical theory of moments. *Ann. Inst. Fourier*, 31(3):99–114, 1981.
- [10] Christian Berg and Marco Thill. Rotation invariant moment problems. *Acta Math.*, 167(3-4):207–227, 1991.

- [11] Horatio Boedihardjo, Xi Geng, Terry Lyons, and Danyu Yang. The signature of a rough path: Uniqueness. *Advances in Mathematics*, 293:720–737, 2016.
- [12] Horatio Boedihardjo, Xi Geng, and Sheng Wang. Cartan’s Path Development, the Logarithmic Signature and a Conjecture of Lyons-Sidorova. *arXiv e-prints: arXiv:2506.18207*, June 2025.
- [13] Richard Brauer. On algebras which are connected with the semisimple continuous groups. *Ann. Math. (2)*, 38:857–872, 1937.
- [14] Mihriban Ceylan and David J. Prömel. Global universal approximation with Brownian signatures. *arXiv e-prints: arXiv:2512.16396*, December 2025.
- [15] Kuo-Tsai Chen. Iterated Integrals and Exponential Homomorphisms†. *Proc. Lond. Math. Soc.*, s3-4(1):502–512, 1954.
- [16] Kuo-Tsai Chen. Integration of paths – a faithful representation of paths by noncommutative formal power series. *Trans. Amer. Math. Soc.*, 89(2):395–407, 1958.
- [17] Ilya Chevyrev, Emilio Ferrucci, Darrick Lee, Terry Lyons, Harald Oberhauser, and Nikolas Tapia. orthsig: Github repository. <https://github.com/emilioferrucci/orthsig>, 2025.
- [18] Ilya Chevyrev, Emilio Ferrucci, Darrick Lee, Terry Lyons, Harald Oberhauser, and Nikolas Tapia. orth: Github repository. <https://github.com/ntapiam/orth>, 2026.
- [19] Ilya Chevyrev and Terry Lyons. Characteristic functions of measures on geometric rough paths. *Ann. Probab.*, 44(6):4049–4082, 2016.
- [20] Ilya Chevyrev and Harald Oberhauser. Signature moments to characterize laws of stochastic processes. *Journal of Machine Learning Research*, 23(176):1–42, 2022.
- [21] Wei-Liang Chow. Über Systeme von linearen partiellen Differentialgleichungen erster Ordnung. *Mathematische Annalen*, 117(1):98–105, December 1940.
- [22] Christa Cuchiero and Janka Möller. Signature methods in stochastic portfolio theory. *SIAM J. Financ. Math.*, 16(4):1239–1303, 2025.
- [23] Christa Cuchiero, Philipp Schmock, and Josef Teichmann. Global universal approximation of functional input maps on weighted spaces. *Constructive Approximation*, 2026.
- [24] Mikael de la Salle (<https://mathoverflow.net/users/10265/mikael-de-la-salle>). Stone-weierstrass analogue for \mathbb{R}^p . MathOverflow. URL:<https://mathoverflow.net/q/96025> (version: 2012-05-04).
- [25] Charles F. Dunkl and Yuan Xu. *Orthogonal Polynomials of Several Variables*. Encyclopedia of Mathematics and Its Applications. Cambridge University Press, Cambridge, 2 edition, 2014.
- [26] Bruno Dupire. Functional Itô calculus. *Quant. Finance*, 19(5):721–729, 2019.
- [27] Bruno Dupire and Valentin Tissot-Daguette. Functional Expansions. *arXiv e-prints: arXiv:2212.13628*, December 2022.
- [28] T. Fawcett. *Problems in stochastic analysis. Connections between rough paths and non-commutative harmonic analysis*. Dphil thesis, University of Oxford, 2003. Supervisor: T. J. Lyons.
- [29] Adeline Fermanian. Embedding and learning with signatures. *Computational Statistics & Data Analysis*, 157:107148, 2021.
- [30] Peter K. Friz and Nicolas B. Victoir. *Multidimensional stochastic processes as rough paths. Theory and applications.*, volume 120 of *Camb. Stud. Adv. Math.* Cambridge: Cambridge University Press, 2010.
- [31] X. Geng and Z. Qian. On an inversion theorem for stratonovich’s signatures of multidimensional diffusion paths. *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, 52(1), February 2016.
- [32] Xi Geng. Reconstruction for the signature of a rough path. *Proc. Lond. Math. Soc. (3)*, 114(3):495–526, 2017.
- [33] Ben Hambly and Terry Lyons. Uniqueness for the signature of a path of bounded variation and the reduced path group. *Ann. Math. (2)*, 171(1):109–167, 2010.
- [34] Michael E. Hoffman. Quasi-shuffle products. *J. Algebr. Comb.*, 11(1):49–68, 2000.
- [35] Yao-Zhong Hu and Paul-André Meyer. Sur les intégrales multiples de stratonovitch. *Séminaire de probabilités de Strasbourg*, 22:72–81, 1988.
- [36] Peter E. Kloeden and Eckhard Platen. *Numerical solution of stochastic differential equations.*, volume 23 of *Appl. Math. (N. Y.)*. Berlin: Springer, 4th corrected printing edition, 2010.
- [37] Yves Le Jan and Zhongmin Qian. Stratonovich’s signatures of brownian motion determine brownian sample paths. *Probability Theory and Related Fields*, 157(1–2):209–223, October 2012.
- [38] G. I. Lehrer and R. B. Zhang. The Brauer category and invariant theory. *J. Eur. Math. Soc. (JEMS)*, 17(9):2311–2351, 2015.
- [39] Siran Li, Zijiu Lyu, Hao Ni, and Jiajie Tao. Restricted Path Characteristic Function Determines the Law of Stochastic Processes. *arXiv e-prints: arXiv:2404.18661*, April 2024.
- [40] Christian Litterer and Harald Oberhauser. On a Chen-Fliess approximation for diffusion functionals. *Monatsh. Math.*, 175(4):577–593, 2014.
- [41] Terry Lyons and Nicolas Victoir. Cubature on Wiener space. *Proc. R. Soc. Lond., Ser. A, Math. Phys. Eng. Sci.*, 460(2041):169–198, 2004.
- [42] Terry J. Lyons. Differential equations driven by rough signals. *Rev. Mat. Iberoam.*, 14(2):215–310, 1998.
- [43] Terry J. Lyons, Michael Caruana, and Thierry Lévy. *Differential equations driven by rough paths. Ecole d’Été de Probabilités de Saint-Flour XXXIV – 2004. Lectures given at the 34th probability summer school, July 6–24, 2004.*, volume 1908 of *Lect. Notes Math.* Berlin: Springer, 2007.
- [44] Terry J. Lyons and Nadia Sidorova. On the radius of convergence of the logarithmic signature. *Illinois Journal of Mathematics*, 50(1-4):763–790, January 2006.
- [45] Edward Nelson. Analytic Vectors. *Annals of Mathematics*, 70(3):572–615, 1959.
- [46] Ivan Nourdin and Giovanni Peccati. *Normal approximations with Malliavin calculus. From Stein’s method to universality*, volume 192 of *Camb. Tracts Math.* Cambridge: Cambridge University Press, 2012.
- [47] Victor Y. Pan. How bad are vandermonde matrices? *SIAM Journal on Matrix Analysis and Applications*, 37(2):676–694, 2016.
- [48] Imanol Perez Arribas. Derivatives pricing using signature payoffs. *arXiv e-prints: arXiv:1809.09466*, September 2018.
- [49] Jeremy Reizenstein. free-lie-algebra-py: Python calculations in tensor space. <https://github.com/bottler/free-lie-algebra-py>, 2025. GitHub repository (accessed 2025-12-25).
- [50] Christophe Reutenauer. *Free Lie Algebras*. London Mathematical Society Monographs. Oxford University Press, Oxford, New York, 1993.
- [51] Jun John Sakurai and Jim Napolitano. *Modern quantum mechanics*. Cambridge: Cambridge University Press, 3rd revised edition edition, 2020.
- [52] Anatolii M. Samoilenko. *Spectral Theory of Families of Self-Adjoint Operators*. Springer Netherlands, 2013.
- [53] R. B. Saxena. Expansion of continuous differentiable functions in Fourier Legendre series. *Can. J. Math.*, 19:823–827, 1967.
- [54] Konrad Schmüdgen. *The Moment Problem*. Springer, November 2017.
- [55] Wim Schoutens. *Stochastic processes and orthogonal polynomials*, volume 146 of *Lect. Notes Stat.* New York, NY: Springer, 2000.
- [56] Daniel W. Stroock and S. R. S. Varadhan. On the support of diffusion processes with applications to the strong maximum principle. *Proc. 6th Berkeley Sympos. math. Statist. Probab., Univ. Calif.* 1970, 3, 333-359 (1972)., 1972.
- [57] Guo-Liang Tian. The comparison between polynomial regression and orthogonal polynomial regression. *Stat. Probab. Lett.*, 38(4):289–294, 1998.

- [58] Anh Tong. signax: Differentiable signature calculations in jax. <https://pypi.org/project/signax/>, 2023. Python package on PyPI, version 0.2.1 (accessed 2025-12-25).
- [59] Lloyd N. Trefethen. *Approximation theory and approximation practice*, volume 164 of *Other Titles Appl. Math.* Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), extended edition edition, 2020.
- [60] Dongbin Xiu and George Em Karniadakis. The Wiener–Askey polynomial chaos for stochastic differential equations. *SIAM J. Sci. Comput.*, 24(2):619–644, 2002.
- [61] Yuan Xu. Unbounded Commuting Operators and Multivariate Orthogonal Polynomials. *Proceedings of the American Mathematical Society*, 119(4):1223–1231, 1993.

APPENDIX A. LINEAR SYSTEMS

We explain the procedure used to study the naturality of the block-orthogonalization map in the shuffle algebra. The full code listings can be found on GitHub [18]. Starting from Fawcett’s formula in (4.2) and the inner product on words in (2.5), we implemented the expected signature and shuffle product in `Wolfram 14`, in order to enable symbolic computations. For simplicity we take $T = 1$. The idea with this approach is that since the natural basis is not orthogonal under (\cdot, \cdot) , inversion of the Gram matrix yielding the projection at degree 5 is computationally intensive.

We begin by creating an abstract symbol w that acts as a placeholder for a word, and declaring it to obey a certain set of rules with respect to the CircleTimes (\otimes) operator, which acts as an external tensor product.

LISTING 1. Tensor product of words

```

1 ClearAll[CircleTimes]
  SetAttributes[CircleTimes, {Flat, OneIdentity}]
3 CircleTimes[x___, y_Plus, z___] := CircleTimes[x, #, z]& /@ y
  CircleTimes[x___, -a_, y___] := -CircleTimes[x, a, y]
5 CircleTimes[x___, a_, y_] := a CircleTimes[x, y] /; FreeQ[a, w]
  CircleTimes[x_, a_, y___] := a CircleTimes[x, y] /; FreeQ[a, w]
7 CircleTimes[x___, a_ y_, z_] := a CircleTimes[x, y, z] /; FreeQ[a, w]
  CircleTimes[x_, a_ y_, z___] := a CircleTimes[x, y, z] /; FreeQ[a, w]
9 CircleTimes[p_ x_] := p CircleTimes[x] /; FreeQ[p, w]

```

We then implement the concatenation operator `Conc`, which maps $T(V) \otimes T(V) \rightarrow T(V)$ linearly:

LISTING 2. Concatenation of words

```

1 ClearAll[Conc]
  SetAttributes[Conc, OneIdentity]
3 Conc[0] = 0;
  Conc[p_ x_] := p Conc[x] /; FreeQ[p, w]
5 Conc[x_Plus] := Plus[Conc/@x]
  Conc[CircleTimes[x_w]] := w@@Delete[0] /@ Level[{x}, 1]
7 Conc[CircleTimes[x_]] := x

```

Finally, the shuffle product on $w[\dots]$ symbols is implemented as a linear map $T(V) \otimes T(V) \rightarrow T(V)$ recursively:

LISTING 3. Shuffle product of words

```

1 ClearAll[Shuf]
  SetAttributes[Shuf, OneIdentity]
3 Shuf[x_w] := x
  Shuf[p_ x_] := p Shuf[x] /; FreeQ[p, w]
5 Shuf[x_Plus] := Plus[Shuf/@x]
  Shuf[x_w\[CircleTimes]w[]] := x
7 Shuf[w[]\[CircleTimes]x_w] := x
  Shuf[w[u_]\[CircleTimes]w[v_]] := Conc[Shuf[w[u]\[CircleTimes]w@@Drop[{v}, -1]]\[
    CircleTimes]w[{v}][[-1]]] + Conc[Shuf[w@@Drop[{u}, -1]\[CircleTimes]w[v]]\[CircleTimes]
    w[{u}][[-1]]]
9 Shuf[x_w\[CircleTimes]y_w] := Shuf[Shuf[CircleTimes[x]]\[CircleTimes]y]

```

The last relation enforces associativity.

Next, we implement (4.2) on pure words $w[\dots]$ and extend by linearity.

LISTING 4. Expected signature of time-augmented Brownian motion

```

1 ClearAll[ESig]
  ESig[w[]]=1;
3 ESig[w[x__]]:=Block[{z=Cases[{x},0],l=DeleteCases[{x},0]},
  With[{n=Length[l],m=Length[z]},
5   If[EvenQ[n],2^(-n/2)/(n/2+m)! Times@@\[Delta]@@@Partition[l,2],0]
  ]
7 ]
  ESig[x_Plus]:=Plus[ESig/@x]
9 ESig[p_ x_]:=p ESig[x]/;FreeQ[p,w]

```

Here, the symbol Δ (δ) is subject to the symmetry rule

```

1 ClearAll[\[Delta]]
  \[Delta]/:\[Delta][b_,a_]:=\[Delta][a,b]/;b>a

```

Finally this induces an inner product on words by (2.5):

LISTING 5. Inner product on the shuffle algebra

```

1 ClearAll[ip]
  ip[x_w,y_w]:=ESig[Shuf[x\[CircleTimes]y]]
3 ip[x_Plus,y_]:=Plus[ip[#,y]&/@x]
  ip[x_,y_Plus]:=Plus[ip[x,#]&/@y]
5 ip[p_ x_,y_]:=p ip[x,y]/;FreeQ[p,w]
  ip[x_,p_ y_]:=p ip[x,y]/;FreeQ[p,w]

```

In the next step, we use these functions to build the Ansatz in Section 4.1, where we also include all crossing partitions.

LISTING 6. Generation of the Ansatz for the orthogonalization map

```

  Pairings[0]=0;Parings[1]=0;
2 Pairings[n_;/;n>1]:=Select[Combinatorica`SetPartitions[n],AllTrue[Length[#]<=2
  &]]
  vars[n_;/;n>0]:=Join[Array[a,Length[Pairings[n]]-1],{1}]
4 Ansatz[n_;/;n>0]:=vars[n].Conc/@Map[CircleTimes@@#&,Map[If[Length[#]==1,w
  @@#,\[Delta]@@#w[]]&,Pairings[n],{2}],{1}]

```

Thus, the command `Ansatz[3]` generates the following output:

$$w[1,2,3] + a[2]w[3]\delta[1,2] + a[3]w[2]\delta[1,3] + a[1]w[1]\delta[2,3]$$

We then generate the orthogonality relations:

```

1 Lower[n_;/;EvenQ[n]]:=With[{pol=Ansatz[n]},Table[ESig[Shuf[pol\[CircleTimes]w@@(n
  +Range[2,k])]],{k,0,n/2-1}]]
  Lower[n_;/;OddQ[n]]:=With[{pol=Ansatz[n]},Table[ESig[Shuf[pol\[CircleTimes]w@@(n
  +Range[2,k+1])]],{k,0,(n-1)/2-1}]]
3  $\delta$ vars[n_]:=DeleteDuplicates[Cases[Lower[n],_ $\delta$ ,Infinity]];

```

The command `Lower[n]` generates a systems of equations that have to be solved for in the a variables. Since these should hold for any choice of letters and not just $w[1,2,3,4]$, we use the `SolveAlways` instruction.

```

1 Sol[n_]:=SolveAlways[#==0&/@Lower[n],\[Delta]vars[n]]

```

For example, running `Sol[3]` yields:

```

1 {{a[1]->-(1/4),a[2]->-(1/4),a[3]->0}}

```

which is precisely the solution in (4.4). Note that $a[3]$, i.e., the variable corresponding to the pairing $\{\{1,3\},\{2\}\}$ is set to zero.

In order to check if the system has a solution, we may look at the matrix corresponding to the system of linear equations generated by `Lower[n]`, say A , and check if the equality $\text{rank}(A) = \text{rank}([A \mid b])$ holds, where b denotes the vector of constant terms. If the equality does not hold, it means that the system is inconsistent and therefore has no solution. In Wolfram this is implemented by the following functions:

```

1 ClearAll[PreCoefs, Coefs, AugCoefs]
  PreCoefs[n_] := PreCoefs[n] = CoefficientArrays[# == 0 & /@ Flatten[
    Values@CoefficientRules[#,  $\delta$ vars[n]] & /@ Lower[n]], Drop[vars[n], -1]] // Normal
3 Coefs[n_] := Coefs[n] = PreCoefs[n][[2]]
  AugCoefs[n_] := Transpose[Insert[Transpose[Coefs[n]], -PreCoefs[n][[1]], -1]]

```

giving $A = \text{Coefs}[n]$ and $[A \mid b] = \text{AugCoefs}[n]$ for a given word length n .

For the case of interest, namely $n = 5$, the matrix has $\text{rank}(A) = 25$ and $\text{rank}([A \mid b]) = 26$. A subset of contradicting equations may be obtained by inspecting this matrix or by computing a certificate by computing a basis of its column null space by solving the system $y^\top A = 0$, and then looking for a vector such that $y^\top b \neq 0$. The nonzero entries of such a vector determine an inconsistent set of equations. This is essentially a version of Farkas' lemma.

In the same way it may be checked that for $n = 6, 7$ the ranks of the corresponding A are 75 and 231, respectively, while the ranks of the augmented matrices are 76 and 232.