Journal of Algebra 634 (2023) 911–938

Contents lists available at ScienceDirect

Journal of Algebra

journal homepage: www.elsevier.com/locate/jalgebra

A structure theorem for streamed information

Cristopher Salvi [a,b,*], Joscha Diehl [c], Terry Lyons [b,d], Rosa Preiss [e], Jeremy Reizenstein [f]

[a] *Imperial College London, United Kingdom of Great Britain and Northern Ireland*
[b] *The Alan Turing Institute, United Kingdom of Great Britain and Northern Ireland*
[c] *University of Greifswald, Germany*
[d] *University of Oxford, United Kingdom of Great Britain and Northern Ireland*
[e] *University of Potsdam, Germany*
[f] *Meta AI, United Kingdom of Great Britain and Northern Ireland*

ARTICLE INFO

Article history:
Received 6 February 2023
Available online 3 August 2023
Communicated by Willem A. de Graaf

Keywords:
Free Lie algebras
Half shuffles
Hall bases
Rough paths
Signatures
Sequential data

ABSTRACT

We identify the free half shuffle algebra of Schützenberger [31] with an algebra of real-valued functionals on paths, where the half shuffle emulates the integration of a functional against another. We then provide two, to our knowledge, new identities in arity 3 involving its commutator (area), and show that these are sufficient to recover the Zinbiel and Tortkara identities introduced by Dzhumadil'daev [11]. We then use these identities to provide a simple proof of the main result of Diehl et al. [8], namely that any element of the free half shuffle algebra can be expressed as a polynomial over iterated areas.

Moreover, we consider minimal sets of Hall iterated integrals defined through the recursive application of the half shuffle product to Hall trees. Leveraging the duality between this set of Hall integrals and classical Hall bases of the free Lie algebra, we prove using combinatorial arguments that any element of the free half shuffle algebra can be written uniquely as a polynomial over Hall integrals. We interpret this result as a structure theorem for streamed information, loosely analogous to the unique prime factorisation of integers, allowing to split any real valued function on streamed data into two parts: a first that extracts and packages the streamed information

[*] Corresponding author.
E-mail address: c.salvi@imperial.ac.uk (C. Salvi).

https://doi.org/10.1016/j.jalgebra.2023.07.024

into recursively defined atomic objects (Hall integrals), and a second that evaluates a polynomial function in these objects without further reference to the original stream. The question of whether a similar result holds if Hall integrals are replaced by Hall areas is left as an open conjecture.
Finally, we construct a canonical, but to our knowledge, new decomposition of the free half shuffle algebra as shuffle power series in the greatest letter of the original alphabet with coefficients in a sub-algebra freely generated by a new alphabet with an infinite number of letters. We use this construction to provide a second proof of our structure theorem.

1. Introduction

It is not too much to accept that, at least on some fine enough time scales, most instance of streamed information (text, sound, video, time series...) can be represented, as a path $\gamma : [0,1] \to V$ with values in some finite dimensional vector space $V \simeq \mathbb{R}^d$. It was first shown by Chen [5], and then explored in greater detail and generality in the context of *rough path theory* in [15,2], that any path may be faithfully represented, up to reparameterisation, by the collection of its iterated integrals known as the *signature*. This non-commutative exponential maps a path to a grouplike element on the tensor algebra $(\mathcal{A}, \otimes)$, where $\mathcal{A}$ is the vector space spanned by words in d letters, including the empty word e, and $\otimes$ is the tensor product. For an arbitrary interval $[a,b] \subset [0,1]$, the signature $\mathcal{S}(\gamma)_{a,b} := X_b$ where X is the unique solution to the control system $dX_t = X_t \otimes d\gamma_t$ started at $X_a = e$. Furthermore, the range of the signature describes the set of characters $G \subset \mathcal{A}$.

The half shuffle product $\prec$ was firstly introduced in [31], where it also showed that $\mathcal{A}$ is the free algebra over A with respect to $\prec$. We will later refer to this algebra as the *free half shuffle algebra of Schützenberger*. In the same article, the shuffle product $\sqcup\!\sqcup$ was subsequently defined as $f \sqcup\!\sqcup g = f \prec g + g \prec f + \langle f, e\rangle\langle g, e\rangle e$, so to emulate integration by parts.

It is well known that the shuffle algebra $(\mathcal{A}, \sqcup\!\sqcup)$ is the algebraic dual of the tensor algebra $(\mathcal{A}, \otimes)$ [28]; it is automatic from this perspective to see that the restriction of linear functionals on $\mathcal{A}$ to the range of the signature G form a unital algebra of real-valued functions that separates points [23]. A straightforward application of the Stone-Weierstrass theorem yields that for any compact set of reparameterisation-reduced paths, linear functionals acting on their signatures are dense in the space of continuous, real-valued functions on this compact set under a suitable choice of topology [4].

Because G is the set of characters, the main result in Ree [26] implies that the restriction of the shuffle product of two of elements of the shuffle algebra to G is the pointwise product of the two restrictions $\langle f \sqcup\!\sqcup g, \mathcal{S}(\gamma)_{a,b}\rangle = \langle f, \mathcal{S}(\gamma)_{a,b}\rangle\langle g, \mathcal{S}(\gamma)_{a,b}\rangle$, the so-called *shuffle identity*. This interplay between algebraic and analytic operations can be extended to the half shuffle product, emulating integration of a path functional against another

$\langle f \prec g, \mathcal{S}(\gamma)_{a,b}\rangle = \int_a^b \langle g, \mathcal{S}(\gamma)_{a,s}\rangle d\langle f, \mathcal{S}(\gamma)_{a,s}\rangle$, and to its commutator representing the area enclosed by the two dimensional curve $t \mapsto (\langle f, \mathcal{S}(\gamma)_{a,t}\rangle, \langle g, \mathcal{S}(\gamma)_{a,t}\rangle)$ and the chord connecting the two end points

$$\langle \mathrm{area}(f,g), \mathcal{S}(\gamma)_{a,b}\rangle = \int_a^b \langle g, \mathcal{S}(\gamma)_{a,s}\rangle d\langle f, \mathcal{S}(\gamma)_{a,s}\rangle - \int_a^b \langle f, \mathcal{S}(\gamma)_{a,s}\rangle d\langle g, \mathcal{S}(\gamma)_{a,s}\rangle.$$

Thus, collectively iterated integrals provide an accurate description of the path and linear combinations of them can be determined easily by regression, making the coefficient of the signature an ideal feature set for machine learning applications on streamed data [13]; signature methods have been applied in a variety of contexts including deep learning for time series [18,25,6], kernel methods [29,21,20] quantitative finance [1,30,16] and cybersecurity [7].

However, these integrals contain some redundancies, in the sense that some higher ones can be expressed using polynomial relations in lower ones. This represents a major scalability issue, particularly because the number of distinct and linearly independent iterated integrals grows exponentially with the degree of iteration in the integral. This raises a simple set of questions which we will answer positively in this paper:

Can we identify minimal sets of integrals so that each integral is an integral of two other integrals in the same class and so that every other integral can be expressed as a polynomial in them?

The minimal sets of integrals we identify in this paper are defined hierarchically using sets of binary planar rooted trees called *Hall sets* [28,3], and can be computed recursively in a localised way (to compute one, one must compute its ancestors but not others) which adds further value to the results. These minimal sets of integrals fully describe the information in the stream while the polynomials capture the nonlinearity in any function of interest. It is for this reason we call it a *structure theorem*, loosely analogous to the unique factorisation of integers as products of primes. In this way we see that identifying a basis for the space of smooth functions acting on pathspace splits the evaluation process into two parts: a) a first that engages with the underlying stream of information,[1] systematically extracts and packages the relevant information into atomic objects whilst removing what's irrelevant, b) a second that evaluates a unique polynomial function in these expensive but informative precomputed basis elements in order to deliver the desired function evaluation without further reference to the original stream γ.

Having established that polynomials in *Hall integrals* freely generate the half shuffle algebra $(\mathcal{A}, \prec)$, it is natural to ask whether a similar structure theorem holds when the

[1] This information extraction is done in practice via some physical integration process that responds to the underlying signal. Physical integration processes are intrinsically nasty as mathematical operators (controlled differential equations in general, and in particular the integration process here, are not closable in the uniform topology on γ - see [24]).

half shuffle $\prec$ is replaced by its commutator area. This question has been, and still remain, a source of conjecture, well supported by calculation, for the last decade. Nonetheless, the search for an answer to this conjecture led us to consider an argument related to the well-known Lazard's elimination [28] to construct a canonical, but to our knowledge, new decomposition of the algebra $\mathcal{A}$ as shuffle power series in the greatest letter of the original alphabet with coefficients in a sub-algebra freely generated by a new alphabet with an infinite number of letters. This construction, that we refer to as elimination trick, will enable us to provide a second proof of our structure theorem relying on an induction argument.

We briefly outline the structure of the paper. Section 2 provides a brief background on the algebraic setup needed for the rest of the paper. In Section 3 we introduce the free half shuffle algebra of Schützenberger, we make precise the interplay between the algebraic operations $\prec, ⧢$, area and the corresponding analytic operations on paths, and we provide two new identities in arity 3 involving the area product. In Section 4 we make use of these new identities to provide a simpler proof of the main result in [8], stating that polynomials in iterated areas generate the algebra $\mathcal{A}$. In Section 5 we present our structure theorem for streamed information, providing a simple proof of the main result in [32] reported without proof also in [17,14] stating that polynomials in Hall integrals freely generate the algebra $\mathcal{A}$. Finally, using the elimination trick we provide a second proof of our structure theorem.

2. Background

First, we remind the reader in a very terse form of the general collection of objects about which we write. Much more can be found by looking in [3] or (and we will follow this for the results we need) [28]. We hope the paper is self contained, and cites what is needed, but for the rest of this introduction, we will be very brief and assume the reader has familiarity with the general algebraic framework.

The starting point will be a finite alphabet A of d letters.

Definition 2.1. A *word* on the alphabet A is a finite sequence of letters from A, including the empty sequence, called the *empty word* and denoted by e. We denote by W_A the set of all words, including the empty word. W_A with the concatenation product is a monoid, that is free over A. The length $|w|$ of a word $w \in W_A$ is the number of letters in w. Finally, we denote by $\mathcal{A}$ the vector space spanned by all words in W_A.

Remark 2.2. The vector space $\mathcal{A}$ admits the unique direct sum decomposition

$$\mathcal{A} = \mathcal{A}^{>0} \oplus \langle e \rangle, \tag{1}$$

where $\langle e \rangle$ is the vector space spanned by the empty word and $\mathcal{A}^{>0}$ is its annihilator, i.e.

$$\mathcal{A}^{>0} := \{f \in \mathcal{A} : \langle f, e \rangle = 0\}.$$

Note that $\mathcal{A}^{>0}$ is the vector space spanned by all non-empty words. It follows that any $f \in \mathcal{A}$ admits the unique decomposition

$$f = (f - \langle f, e\rangle e) + \langle f, e\rangle e,$$

where $(f - \langle f, e\rangle e) \in \mathcal{A}^{>0}$ and $\langle f, e\rangle e \in \langle e\rangle$.

$\mathcal{A}$ is graded by word length. The words of length greater than $n \in \mathbb{N}$ span an ideal, and the quotient of $\mathcal{A}$ by this ideal is often referred to as the *truncated tensor algebra* $\mathcal{A}^{(n)}$.

Definition 2.3. Denote by $(\mathcal{A}, \otimes)$ the *tensor algebra over* A, that is the free associative $\mathbb{R}$-algebra over A with the tensor product $\otimes$.

Remark 2.4. An infinite linear combination of words in W_A is usually referred to as a *series*. There is a natural duality between $\mathcal{A}$ and the associative algebra of all series $\mathcal{A}^\infty$ given by the pairing $(\cdot, \cdot) : \mathcal{A} \times \mathcal{A}^\infty \to \mathbb{R}$ defined as

$$(a, b) = \sum_{\omega \in W_A} a_\omega b_\omega \tag{2}$$

where a_ω, b_ω denote the coefficients in front of the word ω in a, b respectively. Note that this sum is finite because a is a finite linear combination of words. With this pairing, $\mathcal{A}^\infty$ can be identified as the algebraic dual space of $\mathcal{A}$. When restricted to $\mathcal{A} \times \mathcal{A}$, this pairing yields a scalar product with basis W_A and dual basis W_A'. In the sequel we allow implicit and free conversion of letters and words, including the empty word e, according to context use the same notation W_A for the word basis and its dual.

Definition 2.5. The *free magma* $\mathcal{M}_A$ is the minimal non-empty set satisfying: i) $A \subset \mathcal{M}_A$, and ii) if $t', t'' \in \mathcal{M}_A$ then $(t', t'') \in \mathcal{M}_A$. The *degree* of t is defined recursively as $|t| = 1$ if $t \in A$, otherwise if $t', t'' \in \mathcal{M}_A$ then $|t| = |t'| + |t''|$.

Remark 2.6. Let V be a vector space. The space $\mathcal{B}$ of bilinear maps $V \times V \to V$ naturally forms a magma, via composition. For a fixed bilinear map $\phi : V \times V \to V$ and a set map $\iota : A \to V$ we abuse notation and also write $\phi : \mathcal{M}_A \to \mathcal{B}$ for the unique morphism of magmas characterised by

$$\phi(a) = \iota(a), a \in A$$
$$\phi((t', t'')) = \phi(\phi(t'), \phi(t'')).$$

Definition 2.7. The *foliage map* $f : \mathcal{M}_A \to W_A$ is defined on a letter $a \in A$ as $f(a) = a$ and on a tree $t = (t_1, t_2) \in \mathcal{M}_A$ as $f(t) = f(t_1)f(t_2)$ where the product is the tensor product (or concatenation of words).

Remark 2.8. As noted in [28], $\mathcal{M}_A$ can be equivalently identified with the set of binary, planar, rooted trees with leaves labelled in A. For a given element $t \in \mathcal{M}_A$ we will refer to the collection of letters appearing in its leaves as its foliage.

3. The free half shuffle algebra of Schützenberger

In this section we follow [31] to define the half shuffle product and introduce the corresponding free algebra. We also provide two, to our knowledge, new identities in arity 3 involving the commutator of the half shuffle product. These identities will be used in the next section to prove one of the main results of this paper.

Definition 3.1 *([31]).* The *(left) half shuffle product* $\prec : \mathcal{A} \times \mathcal{A} \to \mathcal{A}$ is a bilinear form defined by extending uniquely, by linearity on the decomposition (1), the following relations

1. $e \prec f = 0 \prec f = f \prec 0 = 0$ and $f \prec e = f$, for any $f \in \mathcal{A}^{>0}$, and by induction
2. $f \prec f'' = a(f' \prec f'' + f'' \prec f')$, for any $f = af'$, with $a \in A, f' \in \mathcal{A}^{>0}$, and $f'' \in \mathcal{A}$.

Note that the above definition of $\prec$ is independent of the choice of basis of $\mathcal{A}$.

Remark 3.2. Definition 3.1 differs slightly from the usual algebraic convention that chooses to not define $e \prec e$, as seen e.g. in [12]. In this paper, we follow to the letter [31] where the half shuffle product is defined on $\mathcal{A}^{>0}$ and $\langle e \rangle$, and then extended uniquely to a bilinear map on the direct sum (1) of these two spaces, that is to say the full algebra $\mathcal{A}$. Schützenberger refers to this canonical extension as *prolongment*.

The following theorem is one the main results in [31].

Theorem 3.3. *$\mathcal{A}$ is the free algebra over A with respect to the half shuffle product $\prec$.*

We refer to this algebra as the *free half shuffle algebra of Schützenberger*.

The *shuffle product* $\sqcup\!\sqcup : \mathcal{A} \times \mathcal{A} \to \mathcal{A}$ is defined for any $f, g \in \mathcal{A}$ from the half shuffle $\prec$ as

$$f \sqcup\!\sqcup g = f \prec g + g \prec f + \langle f, e\rangle\langle g, e\rangle e. \tag{3}$$

The algebra $(\mathcal{A}, \sqcup\!\sqcup)$ is an associative and commutative algebra known as the *shuffle algebra*.

Remark 3.4. Note that if $f, g \in \mathcal{A}^{>0}$ then (3) reduces to the more conventional relation

$$f \sqcup\!\sqcup g = f \prec g + g \prec f.$$

The area operator is defined as the commutator of the half shuffle product and will be a core component of the main result in the next section.

Definition 3.5. The operator area : $\mathcal{A} \times \mathcal{A} \to \mathcal{A}$ is the bilinear form defined for $f, g \in \mathcal{A}$ as

$$\mathrm{area}(f,g) = f \prec g - f \prec g. \tag{4}$$

In the next section we will provide concrete examples to demonstrate how Schützenberger's definition of half shuffle is completely consistent with classical integration on paths.

3.1. Schützenberger's half shuffle is consistent with calculus

Consider a smooth path $\gamma : [0,1] \to \mathbb{R}^d$, an interval $[a,b] \subset [0,1]$ and three elements $f, g, h \in \mathcal{A}$.

Define the following one-dimensional paths on $[a,b]$:

$$\mathbf{1} \colon t \mapsto \langle e, \mathcal{S}(\gamma)_{a,t}\rangle, \quad f^\gamma : t \mapsto \langle f, \mathcal{S}(\gamma)_{a,t}\rangle, \quad g^\gamma : t \mapsto \langle g, \mathcal{S}(\gamma)_{a,t}\rangle, \quad h^\gamma : t \mapsto \langle h, \mathcal{S}(\gamma)_{a,t}\rangle.$$

Note that the path $\mathbf{1} \equiv 1$ is constantly equal to 1.

Notice how the relation $e \prec f = 0$ in Definition 3.1 is consistent with the basic fact

$$\langle e \prec f, \mathcal{S}(\gamma)_{a,t}\rangle = \int_a^t f^\gamma_s \, \mathrm{d}\mathbf{1}_s = 0 = \langle 0, \mathcal{S}(\gamma)_{a,t}\rangle,$$

while the relation $f \prec e = f - \langle f, e\rangle e$ is consistent with the fundamental theorem of calculus

$$\langle f \prec e, \mathcal{S}(\gamma)_{a,t}\rangle = \int_a^t \mathbf{1}_s \, \mathrm{d}f^\gamma_s = \int_a^t \mathrm{d}f^\gamma_s = f^\gamma_t - f^\gamma_a = \langle f - \langle f, e\rangle e, \mathcal{S}(\gamma)_{a,t}\rangle \, .$$

All other classical rules of calculus follow. For example integration by parts

$$\langle f \sqcup\!\sqcup g - \langle f, e\rangle \langle g, e\rangle e, \mathcal{S}(\gamma)_{a,t}\rangle = f^\gamma_t g^\gamma_t - f^\gamma_a g^\gamma_a = \int_a^t f^\gamma_s \, \mathrm{d}g^\gamma_s + \int_a^t g^\gamma_s \, \mathrm{d}f^\gamma_s = \langle f \prec g + g \prec f, \mathcal{S}(\gamma)_{a,t}\rangle,$$

follows from the definition of shuffle product in equation (3).

Another classical example is provided by chain rule reads

$$\langle f \prec (g \sqcup\!\sqcup h), \mathcal{S}(\gamma)_{a,t}\rangle = \int_a^t f^\gamma_s g^\gamma_s \, \mathrm{d}h^\gamma_s = \int_a^t f^\gamma_s d \left(\int_a^s g^\gamma_u \, \mathrm{d}h^\gamma_u \right) = \langle (f \prec g) \prec h, \mathcal{S}(\gamma)_{a,t}\rangle,$$

which matches the algebraic relation

$$f \prec (g \sqcup\!\sqcup h) = (f \prec g) \prec h. \tag{5}$$

Equation (5) can be easily verified to hold for letters, and hence for all elements of $\mathcal{A}$ by freeness.

Next we present known and, to our knowledge, new identities on $\mathcal{A}$ involving $\prec, \sqcup\!\sqcup$ and area.

3.2. Identities

The first identity is a direct application of the chain rule and integration by parts. When restricted to $\mathcal{A}^{>0}$ it is known in the literature as *Zinbiel identity* [11].

Lemma 3.6. *For any $f, g, h \in \mathcal{A}$ the following identity holds*

$$(f \prec g) \prec h = f \prec (g \prec h) + f \prec (h \prec g) + \langle g, e\rangle\langle h, e\rangle f \prec e. \tag{6}$$

Proof. A direct application of the chain rule and integration by parts yields

$$\begin{aligned}(f \prec g) \prec h &= f \prec (g \sqcup\!\sqcup h)\\ &= f \prec (g \prec h + h \prec g + \langle g, e\rangle\langle h, e\rangle e)\\ &= f \prec (g \prec h) + f \prec (h \prec g) + \langle g, e\rangle\langle h, e\rangle(f - \langle f, e\rangle e),\end{aligned}$$

and the result follows from equation (3). □

Remark 3.7. When $f, g, h \in \mathcal{A}^{>0}$ equation (6) reduces to the Zinbiel identity

$$(f \prec g) \prec h = f \prec (g \prec h) + f \prec (h \prec g).$$

Remark 3.8. Using Lemma 3.6 it is possible to obtain the following identity

$$f_1 \sqcup\!\sqcup \ldots \sqcup\!\sqcup f_n = \sum_{\sigma \in \mathfrak{S}_n} (\ldots(f_{\sigma(1)} \prec f_{\sigma(2)}) \prec \ldots) \prec f_{\sigma(n)}$$

for any $n \geq 2$ and $f_1, \ldots, f_n \in \mathcal{A}^{>0}$, where $\mathfrak{S}_n$ is the symmetric group of order n.

Remark 3.9. We note an important result obtained by [11] stating that the area operator satisfies no further identity in arity three, but it does satisfy the so-called *Tortkara identity* in arity four. While the Tortkara identity will play no further role in this paper, we mention it here for completeness: for any $f, g, h, i \in \mathcal{A}^{>0}$, we equivalently have

$$\operatorname{area}(\operatorname{area}(f, g), \operatorname{area}(f, h)) = \operatorname{area}(f, \operatorname{vol}(f, g, h))$$

and

$$\begin{aligned}\operatorname{area}(\operatorname{area}(f,g),\operatorname{area}(i,h))+\operatorname{area}(\operatorname{area}(h,g),\operatorname{area}(i,f))&\\=\operatorname{area}(f,\operatorname{vol}(g,h,i))+\operatorname{area}(h,\operatorname{vol}(g,f,i))&\end{aligned}$$

where $\operatorname{vol}(f,g,h):=\operatorname{area}(\operatorname{area}(f,g),h)+\operatorname{area}(\operatorname{area}(g,h),f)+\operatorname{area}(\operatorname{area}(h,f),g)$.

We furthermore note that Tortkara algebras have been studied more in [10], where it has been shown that the span inside $\mathcal{A}$ of iterated areas of letters forms a free Tortkara algebra for $|A|=2$, while the question remains open for larger alphabets.

Remark 3.10 *(Left/right areas).* In this paper, area is defined as the commutator of the left half shuffle. In [8], the right half shuffle is introduced and area is defined as the commutator of the right half shuffle. Although closely connected, these are not identical. The left half shuffle is consistent with [28] and matches the conventions for Hall basis used there (see later sections). The right half shuffle is more consistent with the convention used in integration as the integrand is on the left and the integrator is on the right. The reversed order of terms within equation (5) reflects this dissonance. The proofs of our main results imply equivalent results with the other definition of area, by reversing everything.

Contrary to the Lie bracket $[\cdot,\cdot]$, area does not satisfy the Jacobi identity. However, it satisfies the following two non-trivial and, to our knowledge, new identities that will be leveraged to prove one of the main results of this paper in the next section.

Lemma 3.11 *(Shuffle-pullout identity). For any $f,g,h\in\mathcal{A}$ the following relation holds*

$$\begin{aligned}3\operatorname{area}(h,f\sqcup\!\sqcup g)&=f\sqcup\!\sqcup\operatorname{area}(h,g)+g\sqcup\!\sqcup\operatorname{area}(h,f)-f\sqcup\!\sqcup g\sqcup\!\sqcup h+\langle f,e\rangle\langle g,e\rangle\langle h,e\rangle e\\&\quad+\operatorname{area}(\operatorname{area}(h,g),f)+\operatorname{area}(\operatorname{area}(h,f),g).\end{aligned}$$

Proof. It's easy to check that the relation holds for the empty word e and for letters $a,b,c\in A$

$$\begin{aligned}3\operatorname{area}(c,a\sqcup\!\sqcup b)&=-3\,abc-3\,acb-3\,bac-3\,bca+3\,cab+3\,cba\\&=a\sqcup\!\sqcup\operatorname{area}(c,b)+b\sqcup\!\sqcup\operatorname{area}(c,a)-a\sqcup\!\sqcup b\sqcup\!\sqcup c\\&\quad+\operatorname{area}(\operatorname{area}(c,b),a)+\operatorname{area}(\operatorname{area}(c,a),b).\end{aligned}$$

By Theorem 3.3 we know that $\mathcal{A}$ is free, as a half shuffle algebra over A, therefore the above relation extends to any triple of elements in $\mathcal{A}$. □

Remark 3.12. When $f,g,h\in\mathcal{A}^{>0}$ the shuffle-pullout identity in Lemma 3.11 reduces to

$$3\operatorname{area}(h,f\sqcup\!\sqcup g)=f\sqcup\!\sqcup\operatorname{area}(h,g)+g\sqcup\!\sqcup\operatorname{area}(h,f)-f\sqcup\!\sqcup g\sqcup\!\sqcup h$$

$$+ \operatorname{area}(\operatorname{area}(h, g), f) + \operatorname{area}(\operatorname{area}(h, f), g).$$

Lemma 3.13 *(Area-Jacobi identity). For any triple $f, g, h \in \mathcal{A}$ the following relation is satisfied*

$$\begin{aligned}&\operatorname{area}(\operatorname{area}(f, g), h) + \operatorname{area}(\operatorname{area}(g, h), f) + \operatorname{area}(\operatorname{area}(h, f), g)\\&= -f \sqcup\mathchoice{\mkern-7mu}{\mkern-7mu}{\mkern-3.2mu}{\mkern-3.8mu}\sqcup \operatorname{area}(g, h) - g \sqcup\mathchoice{\mkern-7mu}{\mkern-7mu}{\mkern-3.2mu}{\mkern-3.8mu}\sqcup \operatorname{area}(h, f) - h \sqcup\mathchoice{\mkern-7mu}{\mkern-7mu}{\mkern-3.2mu}{\mkern-3.8mu}\sqcup \operatorname{area}(f, g).\end{aligned}$$

Proof. As before, the relation can be easily verified to hold for e and for letters $a, b, c \in A$:

$$\begin{aligned}&\operatorname{area}(\operatorname{area}(a, b), c) + \operatorname{area}(\operatorname{area}(b, c), a) + \operatorname{area}(\operatorname{area}(c, a), b)\\&= -abc + acb + bac - bca - cab + cba\\&= -a \sqcup\mathchoice{\mkern-7mu}{\mkern-7mu}{\mkern-3.2mu}{\mkern-3.8mu}\sqcup \operatorname{area}(b, c) - b \sqcup\mathchoice{\mkern-7mu}{\mkern-7mu}{\mkern-3.2mu}{\mkern-3.8mu}\sqcup \operatorname{area}(c, a) - c \sqcup\mathchoice{\mkern-7mu}{\mkern-7mu}{\mkern-3.2mu}{\mkern-3.8mu}\sqcup \operatorname{area}(a, b). \quad \square\end{aligned}$$

Remark 3.14. On $\mathcal{A}^{>0}$, starting only from the identities 1) $f \sqcup\mathchoice{\mkern-7mu}{\mkern-7mu}{\mkern-3.2mu}{\mkern-3.8mu}\sqcup g = g \sqcup\mathchoice{\mkern-7mu}{\mkern-7mu}{\mkern-3.2mu}{\mkern-3.8mu}\sqcup f$, 2) $\operatorname{area}(f, g) = -\operatorname{area}(g, f)$, 3) shuffle-pullout, 4) area-Jacobi, it follows from simple calculations that one can recover associativity for $\sqcup\mathchoice{\mkern-7mu}{\mkern-7mu}{\mkern-3.2mu}{\mkern-3.8mu}\sqcup$ and the (left) Zinbiel identity for the left half shuffle $\prec$, now defined by $f \prec g := \frac{1}{2}(f \sqcup\mathchoice{\mkern-7mu}{\mkern-7mu}{\mkern-3.2mu}{\mkern-3.8mu}\sqcup g + \operatorname{area}(f, g))$. Through the Zinbiel identity one then can show the Tortkara identity for $\operatorname{area}(f, g) = f \prec g - g \prec f$ as usual.

4. Polynomials in iterated areas

In this section we present our first main result, namely that polynomial in iterated areas generates the free half-shuffle algebra. We note that this result already appears in [8], however our proof is significantly shorter and based on induction.

4.1. Polynomials in iterated areas are a generating set

Recalling Remark 2.6, we extend area to $\mathcal{M}_A$.

Definition 4.1. $f \in \mathcal{A}$ is an *iterated area* if there exists a tree $t \in \mathcal{M}_A$ so that $f = \mathsf{area}(t)$.

A shuffle monomial of shuffle-degree n is the shuffle product of n iterated areas

$$A_1 \sqcup\mathchoice{\mkern-7mu}{\mkern-7mu}{\mkern-3.2mu}{\mkern-3.8mu}\sqcup \ldots \sqcup\mathchoice{\mkern-7mu}{\mkern-7mu}{\mkern-3.2mu}{\mkern-3.8mu}\sqcup A_n. \tag{7}$$

The empty monomial e has shuffle-degree 0. A shuffle polynomial of shuffle-degree n is a non-degenerate linear combination of such shuffle monomials. Its shuffle-degree is the maximal shuffle-degree of the monomials in the expression.

The sequence defined in the following lemma will play a role in what follows.

Lemma 4.2. *The sequence of negative rationals $\beta_k = -(k-1)/(k+1)$ with $k \geq 1$ is monotone decreasing to -1 and satisfies the following recursion*

$$\beta_1 = 0, \quad \beta_k = \frac{\beta_{k-1} - 1}{\beta_{k-1} + 3}. \tag{8}$$

Exploiting the identities we introduced in the previous section we give a short and direct proof of the main result in [8].

Theorem 4.3. *[8, Corollary 5.6] Any element in $(\mathcal{A}, \prec)$ can be written as a shuffle polynomial in iterated areas $\{\mathsf{area}(t) \mid t \in \mathcal{M}_A\}$.*

Before reproving the theorem we establish the following fundamental re-writing rule that allows one to rewrite the area of a shuffle polynomial in iterated areas with a single iterated area as a new shuffle polynomial in iterated areas, and provides an explicit expression for the monomial of highest shuffle-degree. The proof will crucially depend on both Lemmas 3.11, 3.13.

Theorem 4.4. *For any $n \geq 1$ and any $n+1$ iterated areas $A_1, ..., A_n, A$, the following relation holds*

$$\text{area}(A, A_1 \sqcup\!\sqcup \ldots \sqcup\!\sqcup A_n) = \beta_n A \sqcup\!\sqcup A_1 \sqcup\!\sqcup \ldots \sqcup\!\sqcup A_n + Q, \tag{9}$$

where $\beta_n = -(n-1)/(n+1)$, and Q is a shuffle polynomial in iterated areas of shuffle-degree at most n.

Remark 4.5. Note that it remains an open problem whether $\alpha = \beta_n$ is the only real number such that

$$\text{area}(a, a_1 \sqcup\!\sqcup \ldots \sqcup\!\sqcup a_n) - \alpha a \sqcup\!\sqcup a_1 \sqcup\!\sqcup \ldots \sqcup\!\sqcup a_n$$

can be expressed as a shuffle polynomial in iterated areas of shuffle-degree at most n for any letters $a, a_1, \ldots, a_n$. This question arises due to the fact that iterated areas do not *freely* generate the shuffle algebra. However, for the example $n = 2$, $\beta_2 = -1/3$ is indeed the only such coefficient because the area-Jacobi identity is the only relation between iterated areas on level 3.

Proof. We prove the statement (9) by induction on n. If $n = 1$ then the statement is trivially true, with $\beta_1 = 0$ and $Q = \text{area}(A_1, A)$.

Suppose the statement (9) holds for any $n < k$. Consider k iterated areas $A_1, ..., A_k$ and an additional iterated area A. We recall that the shuffle product $\sqcup\!\sqcup$ is associative and commutative on $\mathcal{A}$. By the shuffle-pullout identity we have

$$\begin{aligned}3\,\mathrm{area}(A, A_1 \sqcup\!\sqcup \dots \sqcup\!\sqcup A_k) &= A_1 \sqcup\!\sqcup \mathrm{area}(A, A_2 \sqcup\!\sqcup \dots \sqcup\!\sqcup A_k)\\ &\quad + A_2 \sqcup\!\sqcup \dots \sqcup\!\sqcup A_k \sqcup\!\sqcup \mathrm{area}(A, A_1)\\ &\quad - A_1 \sqcup\!\sqcup \dots \sqcup\!\sqcup A_k \sqcup\!\sqcup A\\ &\quad + \mathrm{area}(\mathrm{area}(A, A_1), A_2 \sqcup\!\sqcup \dots \sqcup\!\sqcup A_k)\\ &\quad + \mathrm{area}(\mathrm{area}(A, A_2 \sqcup\!\sqcup \dots \sqcup\!\sqcup A_k), A_1).\end{aligned}$$

By induction ($n = k - 1$) we have that

$$\begin{aligned}A_1 \sqcup\!\sqcup \mathrm{area}(A, A_2 \sqcup\!\sqcup \dots \sqcup\!\sqcup A_k) &= A_1 \sqcup\!\sqcup (\beta_{k-1} A \sqcup\!\sqcup A_2 \sqcup\!\sqcup \dots \sqcup\!\sqcup A_k + Q_1')\\ &= \beta_{k-1} A \sqcup\!\sqcup A_1 \sqcup\!\sqcup \dots \sqcup\!\sqcup A_k + A_1 \sqcup\!\sqcup Q_1'\end{aligned}$$

where $A_1 \sqcup\!\sqcup Q_1'$ is a shuffle-polynomial of shuffle-degree k. By definition $\mathrm{area}(A, A_1)$ is a iterated area and so

$$Q_2' = A_2 \sqcup\!\sqcup \dots \sqcup\!\sqcup A_k \sqcup\!\sqcup \mathrm{area}(A, A_1)$$

is a shuffle monomial of shuffle-degree k. Similarly, the induction hypothesis implies that

$$Q_3' = \mathrm{area}(\mathrm{area}(A, A_1), A_2 \sqcup\!\sqcup \dots \sqcup\!\sqcup A_k)$$

is a shuffle-polynomial of shuffle-degree k, where $\hat{Q}_3'$ is a shuffle-polynomial of shuffle-degree $k - 1$. Hence, $Q' = A_k \sqcup\!\sqcup Q_1' + Q_2' + Q_3'$ is a shuffle polynomial of shuffle-degree k and

$$\begin{aligned}3\,\mathrm{area}(A, A_1 \sqcup\!\sqcup \dots \sqcup\!\sqcup A_k) &= Q' + (\beta_{k-1} - 1) A_1 \sqcup\!\sqcup \dots \sqcup\!\sqcup A_k \sqcup\!\sqcup A\\ &\quad + \mathrm{area}(\mathrm{area}(A, A_2 \sqcup\!\sqcup \dots \sqcup\!\sqcup A_k), A_1).\end{aligned} \tag{10}$$

It remains to consider the last term $\mathrm{area}(\mathrm{area}(A, A_2 \sqcup\!\sqcup \dots \sqcup\!\sqcup A_k), A_1)$.

By the area-Jacobi identity and the anticommutativity of area we can rewrite this term as follows

$$\begin{aligned}\mathrm{area}(\mathrm{area}(A, A_2 \sqcup\!\sqcup \dots \sqcup\!\sqcup A_k), A_1) &= \mathrm{area}(\mathrm{area}(A_1, A_2 \sqcup\!\sqcup \dots \sqcup\!\sqcup A_k), A)\\ &\quad - \mathrm{area}(\mathrm{area}(A_1, A), A_2 \sqcup\!\sqcup \dots \sqcup\!\sqcup A_k)\\ &\quad + A \sqcup\!\sqcup \mathrm{area}(A_1, A_2 \sqcup\!\sqcup \dots \sqcup\!\sqcup A_k)\\ &\quad - A_2 \sqcup\!\sqcup \dots \sqcup\!\sqcup A_k \sqcup\!\sqcup \mathrm{area}(A_1, A)\\ &\quad - A_1 \sqcup\!\sqcup \mathrm{area}(A, A_2 \sqcup\!\sqcup \dots \sqcup\!\sqcup A_k).\end{aligned}$$

Again, $\mathrm{area}(A_1, A)$ is a iterated area, and by induction the term

$$Q_1'' = -\,\mathrm{area}(\mathrm{area}(A_1, A), A_2 \sqcup\!\sqcup \dots \sqcup\!\sqcup A_k)$$

is a polynomial in iterated areas of shuffle-degree at most k. The term

$$Q_2'' = -A_2 \sqcup\!\sqcup \ldots \sqcup\!\sqcup A_k \sqcup\!\sqcup \operatorname{area}(A_1, A)$$

is clearly a monomial in iterated areas of shuffle-degree k. By induction we have that

$$P_1 = \operatorname{area}(A_1, A_2 \sqcup\!\sqcup \ldots \sqcup\!\sqcup A_k) = \beta_{k-1} A_1 \sqcup\!\sqcup \ldots \sqcup\!\sqcup A_k + P_1'$$

where P_1' is a polynomial in iterated areas of shuffle-degree $k-1$. Similarly

$$P_2 = \operatorname{area}(A, A_2 \sqcup\!\sqcup \ldots \sqcup\!\sqcup A_k) = \beta_{k-1} A \sqcup\!\sqcup A_2 \sqcup\!\sqcup \ldots \sqcup\!\sqcup A_k + P_2'$$

where P_2' is a polynomial in iterated areas of shuffle-degree $k-1$. Therefore

$$Q_3'' = A \sqcup\!\sqcup \operatorname{area}(A_1, A_2 \sqcup\!\sqcup \ldots \sqcup\!\sqcup A_k) = \beta_{k-1} A \sqcup\!\sqcup A_1 \sqcup\!\sqcup \ldots A_k + A \sqcup\!\sqcup P_1'.$$

Similarly

$$Q_4'' = -A_1 \sqcup\!\sqcup \operatorname{area}(A, A_2 \sqcup\!\sqcup \ldots \sqcup\!\sqcup A_k) = -\beta_{k-1} A \sqcup\!\sqcup A_1 \sqcup\!\sqcup \ldots A_k - A_1 \sqcup\!\sqcup P_2'.$$

Combining terms we get a cancellation and degree reduction so that

$$\begin{aligned} Q_3'' + Q_4'' &= \beta_{k-1} A_1 \sqcup\!\sqcup \ldots A_k \sqcup\!\sqcup A + A \sqcup\!\sqcup P_1' - \beta_{k-1} A_1 \sqcup\!\sqcup \ldots A_k \sqcup\!\sqcup A - A_1 \sqcup\!\sqcup P_2' \\ &= A \sqcup\!\sqcup P_1' - A_1 \sqcup\!\sqcup P_2' \end{aligned}$$

is a polynomial in iterated areas of shuffle-degree k. Setting $Q'' = Q_1'' + Q_2'' + Q_3'' + Q_4''$ (which is a polynomial in iterated areas of shuffle degree k) and substituting in equation (10) we get

$$\begin{aligned} 3 \operatorname{area}(A, A_1 \sqcup\!\sqcup \ldots \sqcup\!\sqcup A_k) &= (\beta_{k-1} - 1) A_1 \sqcup\!\sqcup \ldots \sqcup\!\sqcup A_k \sqcup\!\sqcup A \\ &\quad + \operatorname{area}(\operatorname{area}(A_1, A_2 \sqcup\!\sqcup \ldots \sqcup\!\sqcup A_k), A) + Q' + Q'' \\ &= (\beta_{k-1} - 1) A_1 \sqcup\!\sqcup \ldots \sqcup\!\sqcup A_k \sqcup\!\sqcup A + \operatorname{area}(P_1, A) + Q' + Q'' \end{aligned} \tag{11}$$

P_1 being a polynomial in iterated areas of shuffle-degree $k-1$, we have by induction that $\operatorname{area}(P_1, A)$ is a polynomial in iterated areas of shuffle-degree k. Hence, by construction $Q = Q' + Q'' + \operatorname{area}(P_1, A)$ is a polynomial in iterated areas of shuffle-degree k. Therefore equation (11) becomes

$$\begin{aligned} 3 \operatorname{area}(A, A_1 \sqcup\!\sqcup \ldots \sqcup\!\sqcup A_k) &= (\beta_{k-1} - 1) A_1 \sqcup\!\sqcup \ldots \sqcup\!\sqcup A_k \sqcup\!\sqcup A \\ &\quad - \beta_{k-1} \operatorname{area}(A, A_1 \sqcup\!\sqcup \ldots \sqcup\!\sqcup A_k) + Q. \end{aligned}$$

Rearranging the terms we get the following final expression

$$\mathrm{area}(A, A_1 \sqcup\!\sqcup \ldots \sqcup\!\sqcup A_k) = \frac{\beta_{k-1} - 1}{\beta_{k-1} + 3} A_1 \sqcup\!\sqcup \ldots \sqcup\!\sqcup A_k \sqcup\!\sqcup A + \frac{1}{\beta_{k-1} + 3} Q.$$

Setting $\beta_k = \frac{\beta_{k-1}-1}{\beta_{k-1}+3}$ and noting that $\beta_1 = 0$ the result follows from Lemma 4.2. □

Proof of Theorem 4.3. Since linear combinations of polynomials are polynomials, it suffices to prove that words in W_A are polynomial in iterated areas. We prove by induction that every word $w \in W_A$ of length $|w| = n$ can be expressed as polynomial in iterated areas of shuffle-degree n. The result is trivial for $n = 0$. Let $n \geq 1$. We assume that w is a word of length $n > 0$ and that any word of length $< n$ can be written as a polynomial in iterated areas of the appropriate degree.

Since $|w| > 0$, w can be written as follows

$$w = av = a \prec v \tag{12}$$

where $v \in W_A$ is of word of length $|v| = n - 1$ and $a \in A \subset \mathcal{A}$ is a letter. Moreover for any elements of $\mathcal{A}$

$$\begin{aligned} a \prec v &= \frac{1}{2}(\mathrm{area}(a, v) + a \sqcup\!\sqcup v - (a, e)(v, e)e) && (13) \\ &= \frac{1}{2}(\mathrm{area}(a, v) + a \sqcup\!\sqcup v) && (14) \end{aligned}$$

since a is a letter. The length of the word v in (13) is equal to $n - 1$, so by induction it can be written as a polynomial in iterated areas of shuffle-degree $n - 1$. Hence, the term $a \sqcup\!\sqcup v$ is a shuffle polynomial in iterated areas of shuffle-degree n. By Theorem 4.4 the term $\mathrm{area}(a, v)$ is also a polynomial in iterated areas of shuffle-degree n, and so w a polynomial in iterated areas of shuffle-degree n. This concludes the induction and the proof. □

5. A structure theorem for streamed information

To present our structure theorem we will need to introduce the free Lie algebra $\mathcal{L}_A$ over A.

5.1. The free Lie algebra

$(\mathcal{A}, [\cdot, \cdot])$ is also a Lie algebra with Lie bracket $[x, y] = x \otimes y - y \otimes x$ for $x, y \in \mathcal{A}$.

Definition 5.1. Denote by $(\mathcal{L}_A, [\cdot, \cdot])$ the Lie algebra generated by A in $\mathcal{A}$, i.e. the intersection of all Lie algebras in $\mathcal{A}$ containing A.

Lemma 5.2. *[28, Theorem 0.5]* $(\mathcal{L}_A, [\cdot,\cdot])$ *is the free Lie algebra over* A.

Remark 5.3. The maps exp and log are classically defined as power series mapping $\mathcal{A}^\infty$ to $\mathcal{A}^\infty$. The truncated power series for $\exp^{(n)}$ and $\log^{(n)}$ provide good meaning for these operators as maps from $\mathcal{A}$ into $\mathcal{A}$. Those elements in $\mathcal{A}$ that are, at each truncated level $n \in \mathbb{N}$, in $\mathcal{L}_A$ are known as *Lie elements* and denoted by $\mathcal{L}_A^{(n)}$. Those elements in $\mathcal{A}$ that are, at each truncated level, exponentials of Lie elements, or equivalently, whose truncated logarithm is in $\mathcal{L}_A$, are known as *grouplike elements* (and they form a group). The maps log and exp provide a one to one correspondence between group-like elements and Lie elements.

We report the following three classical results about the shuffle product $\sqcup\!\sqcup$ and the free Lie algebra $\mathcal{L}_A$: the first states that the shuffle product characterises grouplike elements [23, Lemma 2.17], the second provides a characterisation of Lie elements in $\mathcal{L}_A$ [28, Theorem 3.1 (iv)], and the third states that the exponential of Lie elements span the tensor algebra in a way that respects degrees of truncation [9, Lemma 3.4].

Theorem 5.4. *Let* $\ell \in \mathcal{L}_A$ *be a Lie element.*

1. $\langle f, \exp(\ell)\rangle\langle g, \exp(\ell)\rangle = \langle f \sqcup\!\sqcup g, \exp(\ell)\rangle$ *for any* $f, g \in \mathcal{A}$.
2. $\langle f \sqcup\!\sqcup g, \ell\rangle = 0$ *for any* $f, g \in \mathcal{A}^{>0}$.
3. $\mathcal{A}^{(n)} = Span\{\exp^{(n)}(\ell) : \ell \in \mathcal{L}_A^{(n)}\}$ *for any degree of truncation* $n \in \mathbb{N}$.

In light of Theorem 5.4 and of the following Lemma, the shuffle algebra $(\mathcal{A}, \sqcup\!\sqcup)$ can be identified with the algebra of $\mathbb{Q}$-polynomial functions on $\mathcal{L}_A$ with pointwise multiplication, denoted by $\mathbb{Q}[\mathcal{L}_A]$.

Lemma 5.5. *For any* $f \in \mathcal{A}$, *the map* $\ell \mapsto \langle f, \exp(\ell)\rangle$ *is in* $\mathbb{Q}[\mathcal{L}_A]$. *Furthermore, the map* $f \mapsto \langle f, \exp(\cdot)\rangle$ *from* $\mathcal{A}$ *to* $\mathbb{Q}[\mathcal{L}_A]$ *is bijective.*

Proof. This result is classical, so we provide only a sketch of the proof. Any element $x \in \mathcal{A}$ is a finite sum of words in W_A of some maximal length $d(x)$. Fix some basis $(\ell_i)_i$ for $\mathcal{L}_A$ that respects dimension and let $\ell = \sum l_i \ell_i$. Then the map $(s, \exp(\ell)) = (s, \exp(\sum_{d(\ell_i)\leq d(x)} l_i \ell_i))$ and the right hand side, truncated at degree $d(x)$ is clearly a polynomial in the l_i. The exponentials of truncated Lie elements are linearly dense in the truncated tensor algebra, therefore x is completely determined by its inner product with the $(x, \exp(\ell))$ as ℓ varies. □

Remark 5.6. It is an immediate corollary of these results, and of the *Stone Weierstrass Theorem*, that any finite collection of distinct grouplike elements form the vertices of a simplex, and therefore that there is a linear functional that is one on any one of the elements and zero on the others.

Remark 5.7. An analogy can be drawn with the Fourier transform seen as a change of basis for signals from time to frequency domain that turns point-wise multiplication into convolution. In our case, we can view $(\mathcal{A}, \sqcup\!\sqcup)$ as polynomial functions on $\mathcal{L}_A$ with pointwise multiplication, or as an algebra spanned by words, with the shuffle product, depending on our viewpoint.

Next we introduce special subsets of Hall trees in $\mathcal{M}_A$ classically used to construct bases for $\mathcal{L}_A$. Recall Remark 2.6 stating that any binary operator defined on words over A automatically extends to an operator acting on trees from the magma $\mathcal{M}_A$. In particular, this extends the Lie bracket, the half shuffle $\prec$, and the operation area, to maps from $\mathcal{M}_A$ to $\mathcal{A}$.

5.2. Hall sets

Definition 5.8. A total order $<$ on a subset M of $\mathcal{M}_A$ is an *ancestral order* if for any tree $t = (t', t'')$ of degree ≥ 2 one has $t < t''$.

This definition of ancestral order makes other constructions more transparent. It is obvious that ancestral orders exist on any magma and their restrictions to a subset are also ancestral.

Definition 5.9. A subset H of $\mathcal{M}_A$ together with an order $<$ on H is a *Hall set* if the following conditions hold

1. $<$ is an ancestral order on H;
2. $A \subset H$;
3. for any tree $h = (h_1, h_2) \in \mathcal{M}_A$ of degree ≥ 2, $h \in H$ if and only if:
   (a) $h_1, h_2 \in H$ and $h_1 < h_2$
   (b) either $h_1 \in A$ or $h_2 \leq h_1''$ where $h_1 = (h_1', h_1'')$.

We note that, since $<$ is assumed to be ancestral, point 3.a implies $h < h_2$, which is a condition needed in the general definition of Hall sets. As pointed out in [28, Proposition 4.1] and the surrounding discussion, Hall sets exist, any ancestral order on the full magma leads in a canonical way to a unique Hall set, and that Hall sets are *closed*, i.e. each subtree of a Hall tree is again a Hall tree.

Example 5.10. The Hall set H set used in the `esig` package [22] is defined as follows: elements are ordered so that they respect degree, and for any equal-length Hall trees $h = (h_1, h_2), h' = (h_1', h_2')$ their order is defined recursively as follows: $h < h'$ if either $h_1 < h_1'$ or $h_1 = h_1'$ and $h_2 < h_2'$.

Example 5.11. Consider a total order on letters in A and suppose that words in W_A are ordered alphabetically. A *Lyndon word* on W_A is a non-empty word such that for any factorisation $\omega = uv$ with $u, v \in W_A$ non-empty one has $\omega < v$. Then, the set of Lyndon words ordered alphabetically is a Hall set [28, Theorem 5.1].

Example 5.12. Let $H_0 = A$ and order it totally. Define H_{n+1} as the set of trees of the form

$$h = (...((h_1, h_2), h_3), ..., h_k)$$

where $k \geq 2$ and $h_1, ..., h_k \in H_n$ with

$$h_1 < h_2 \geq h_3 \geq ... \geq h_k.$$

Now order H_{n+1} totally. Finally let $H = \cup_{n \geq 0} H_n$ and extend the order in H_n to H by the condition

$$h_1 = H_m, h_2 \in H_n, m < n \implies h_1 > h_2.$$

Then H is a Hall set [28, Theorem 5.7].

Lemma 5.13. *[28, Corollary 4.14] Let A be an alphabet of q letters. The number of Hall trees of degree n is equal to*

$$\mathcal{D}_H = \frac{1}{n} \sum_{d|n} \mu(d) q^{n/d} \tag{15}$$

where μ is the Möbius function.

5.3. The Poincaré-Birkhoff-Witt basis and its dual

The Jacobi identities are linear relations between degree-three Lie brackets arising from associativity of the underlying group operation. They make the derivation of a basis for the free Lie algebra $\mathcal{L}_A$ a deep and classic challenge.

Theorem 5.14. *[28, Theorem 4.9 (i)] For any Hall set H, the collection of elements $\{[h] : h \in H\}$ form a linear basis for the free Lie algebra $\mathcal{L}_A$.*

This basis admits a canonical extension to a basis of the tensor algebra $(\mathcal{A}, \otimes)$.

Theorem 5.15. *[28, Theorem 4.9] The decreasing products*

$$[h_1]^{\otimes k_1} \otimes ... \otimes [h_n]^{\otimes k_n}, \quad h_i \in H, \quad h_1 > ... > h_n \tag{16}$$

are a basis of the tensor algebra $(\mathcal{A}, \otimes)$. This basis is called the Poincaré-Birkhoff-Witt (PBW) basis.

Definition 5.16. A word $\omega \in W_A$ is called a Hall word if ω is the image of a Hall tree $h \in H$ by the foliage map, i.e. $\omega = f(h)$.

Remark 5.17. The foliage map is injective when restricted to a Hall set H and there are efficient algorithms for recovering the Hall tree from a Hall word.

Lemma 5.18. *[28, Corollary 4.7] Every word $\omega \in W_A$ can be written uniquely as a decreasing product of Hall words*

$$\omega = f(h_1)^{\otimes k_1} \otimes \ldots \otimes f(h_n)^{\otimes k_n}, \quad h_i \in H, \quad h_1 > \ldots > h_n. \tag{17}$$

Remark 5.19. If $\omega \in W_A$ is a word decomposed into its unique decreasing product of Hall words according to equation (17), then P_ω is the corresponding PBW basis element as per Theorem 5.15

$$P_\omega = [h_1]^{\otimes k_1} \otimes \ldots \otimes [h_n]^{\otimes k_n}, \quad h_i \in H, \quad h_1 > \ldots > h_n.$$

$\{P_\omega\}_{\omega \in W_A}$ is thus an enumeration of the PBW basis indexed by words. The next theorem provides exact formulae for the dual basis to the PBW basis.

Theorem 5.20. *[28, Theorem 5.3] The dual basis $\{S_\omega\}_{\omega \in W_A}$ to the PBW basis $\{P_\omega\}_{\omega \in W_A}$ has the following properties:*

1. *If e is the empty word then $S_e = e$.*
2. *If $\omega = f(h_1)^{\otimes k_1} \otimes \ldots \otimes f(h_n)^{\otimes k_n}$ is the unique factorisation of the word ω in a decreasing product of Hall trees $h_1 > \ldots > h_n \in H$, then*

$$S_\omega = \frac{1}{k_1! \ldots k_n!} S_{f(h_1)}^{\sqcup\!\sqcup k_1} \sqcup\!\sqcup \ldots \sqcup\!\sqcup S_{f(h_n)}^{\sqcup\!\sqcup k_n}. \tag{18}$$

3. *If $h \in H$, then the word $f(h) = av$ for some letter $a \in A$ and word $v \in W_A$; moreover*

$$S_{f(h)} = a \otimes S_v. \tag{19}$$

Theorem 5.20 is an important result due to Schützenberger and it is the structure theorem mentioned in the introduction. However, in the next section we provide our version of this theorem (which agrees with the version in [32] but with a completely different proof) which consists of a more explicit recursive formula for the dual PBW basis elements $\{S_\omega\}_{\omega \in W_A}$ and identify them as Hall integrals. We note that this result is reported without proof also in [17,14].

5.4. Polynomials in Hall integrals are a free generating set

Definition 5.21. An element x of $\mathcal{A}$ is called a *Hall integral* if it is the image under the operator $\prec : \mathcal{M}_A \to \mathcal{A}$ of a Hall tree. That is to say, there exists a Hall tree $h \in H \subset \mathcal{M}_A$

so that $x = \prec(h)$. A (shuffle) polynomial in Hall integrals is a sum of shuffle monomials in Hall integrals.

The following Lemma follows immediately from the definition of a Hall tree.

Lemma 5.22. *Any Hall tree $h \in H$ can be uniquely decomposed as*

$$h = (h_1 h_2^k) = (...((h_1, h_2), h_2), ...h_2) \tag{20}$$

with $h_1, h_2 \in H$, and either h_1 is a letter or $h_1'' \neq h_2$ and where the h_2 bracketing is repeated k times. This is often referred to as the Lazard decomposition *of h.*

Definition 5.23. If $h = (h_1 h_2^k)$ is the Lazard decomposition of a Hall tree $h \in H$ then we define the *Lazard depth* α_h of h to be $1/k$. The *accumulated Lazard depth* of a Hall tree $h \in H$ is defined recursively: $\mathcal{A}_h = 1$ if $h \in A$, otherwise $h = (h', h'')$ and $\mathcal{A}_h = \alpha_h \mathcal{A}_{h'} \mathcal{A}_{h''}$.

The following are the main results of this section.

Theorem 5.24. *For any Hall tree $h \in H \setminus A$ one has $h = (h', h'')$ and*

$$S_{f(h)} = \alpha_h \left(S_{f(h')} \prec S_{f(h'')} \right) \tag{21}$$

where $\alpha_h \in \mathbb{Q}$ is the Lazard depth of h.

Theorem 5.25. *For any Hall tree $h \in H$ one has*

$$S_{f(h)} = \mathcal{A}_h(\prec(h)) \tag{22}$$

where $\mathcal{A}_h \in \mathbb{Q}$ is the accumulated Lazard depth of h.

Theorem 5.26. *Consider all decreasing sequences $h_i \in H$, $h_1 > ... > h_n$, and strictly positive integers $k_i > 0$; then the elements*

$$S_\omega = \frac{\mathcal{A}_{h_1}^{k_1} \dots \mathcal{A}_{h_n}^{k_n}}{k_1! ... k_n!} (\prec(h_1))^{\sqcup\!\sqcup k_1} \sqcup\!\sqcup ... \sqcup\!\sqcup (\prec(h_n))^{\sqcup\!\sqcup k_n} \tag{23}$$

are the dual basis in $\mathcal{A}$ to the PBW basis $\{P_\omega = [h_1]^{\otimes k_1} \otimes ... \otimes [h_n]^{\otimes k_n}\}_{\omega \in W_A}$.

Before proving Theorem 5.24 we need the following combinatorial lemma.

Lemma 5.27. *[28, Corollary 5.14] Let $h = (h', h'') \in H$ be a Hall tree. Now $f(h) = av$, where $a \in A$ and $v \in W_A$. Let $v = f(h_1)^{\otimes k_1} \otimes ... \otimes f(h_n)^{\otimes k_n}$ be the unique factorisation of the word v in a decreasing product of Hall trees $h_1 > ... > h_n \in H$. Then*

$$h'' = h_n. \tag{24}$$

Proof of Theorem 5.24. We write $f(h) = av$, with $a \in A$ and $v \in W_A$. Let $v = f(h_1)^{\otimes k_1} \otimes ... \otimes f(h_n)^{\otimes k_n}$ be the unique factorisation of the word v in a decreasing product of Hall trees $h_1 > ... > h_n \in H$. By Lemma 5.27 $h'' = h_n$. By Theorem 5.20 we also know that

$$\begin{aligned}
S_{f(h)} &= a \otimes S_v &(25)\\
&= S_a \prec S_v &(26)\\
&= \frac{1}{k_1!...k_n!} S_a \prec (S_{f(h_1)}^{\sqcup\!\sqcup k_1} \sqcup\!\sqcup ... \sqcup\!\sqcup S_{f(h_n)}^{\sqcup\!\sqcup k_n}) &(27)\\
&= \frac{1}{k_1!...k_n!} S_a \prec ((S_{f(h_1)}^{\sqcup\!\sqcup k_1} \sqcup\!\sqcup ... \sqcup\!\sqcup S_{f(h_n)}^{\sqcup\!\sqcup k_n - 1}) \sqcup\!\sqcup S_{f(h_n)}) &(28)\\
&= \frac{1}{k_1!...k_n!} S_a \prec ((S_{f(h_1)}^{\sqcup\!\sqcup k_1} \sqcup\!\sqcup ... \sqcup\!\sqcup S_{f(h_n)}^{\sqcup\!\sqcup k_n - 1}) \sqcup\!\sqcup S_{f(h'')}) &(29)\\
&= \frac{1}{k_1!...k_n!} (S_a \prec (S_{f(h_1)}^{\sqcup\!\sqcup k_1} \sqcup\!\sqcup ... \sqcup\!\sqcup S_{f(h_n)}^{\sqcup\!\sqcup k_n - 1})) \prec S_{f(h'')}. &(30)
\end{aligned}$$

Equation (25) is a restatement of (19) in Theorem 5.20. Equation (26) is immediate from the definition of $\prec$. Equation (27) follows from (37) in Theorem 5.20. Equation (28) is simply the associative property of shuffle. Equation (29) follows from Lemma 5.27. Equation (30) follows from the chain rule (5). Note that the inner term in equation (30) can be reinterpreted as $S_{f(h')}$ (up to scalar) because

$$S_{f(h')} = \frac{1}{k_1!...(k_n - 1)!} S_a \prec (S_{f(h_1)}^{\sqcup\!\sqcup k_1} \sqcup\!\sqcup ... \sqcup\!\sqcup S_{f(h_n)}^{\sqcup\!\sqcup k_n - 1}). \tag{31}$$

Substituting this into equation (30) and recalling the definition of the Lazard depth α_h we obtain

$$\begin{aligned}
S_{f(h)} &= \frac{1}{k_n} (S_{f(h')} \prec S_{f(h'')}) &(32)\\
&= \alpha_h (S_{f(h')} \prec S_{f(h'')}). \quad \square &(33)
\end{aligned}$$

Proof of Theorem 5.25. We may proceed by induction. For any Hall tree $h \in A$ one has $S_{f(h)} = h \in \mathcal{A}$, $\prec(h) = h$, and $\mathcal{A}_h = 1$ and so the theorem is true. On the other hand if $h = (h', h'')$ then, assuming the result holds for h', h'':

$$\begin{aligned}
S_{f(h)} &= \alpha_h \left(S_{f(h')} \prec S_{f(h'')} \right) &(34)\\
&= \alpha_h \left((\mathcal{A}_{h'}(\prec(h'))) \prec (\mathcal{A}_{h''}(\prec(h''))) \right) &(35)\\
&= \mathcal{A}_h(\prec(h)) &(36)
\end{aligned}$$

where we use Theorem 5.24 for the first step, the truth of the result for h' and h'' for the second, and the recursive definitions of $\prec(h)$ and $\mathcal{A}_h$ for the third. So the result is true for h. □

Proof of Theorem 5.26. Recall from Schützenberger's theorem (Theorem 5.20 in this paper) that any element S_w in the dual basis to the PWB basis can be expressed uniquely as a shuffle monomial in $S_{f(h)}$. More precisely, consider the unique factorisation of the word w as a decreasing product of Hall words $w = f(h_1)^{\otimes k_1} \otimes ... \otimes f(h_n)^{\otimes k_n}$ where $h_1 > ... > h_n \in H$, then the dual basis element

$$S_w = \frac{1}{k_1!...k_n!} S_{f(h_1)}^{\sqcup\!\sqcup k_1} \sqcup\!\sqcup ... \sqcup\!\sqcup S_{f(h_n)}^{\sqcup\!\sqcup k_n} \in \mathcal{A}. \tag{37}$$

Theorem 5.25 allows for $i = 1...n$ the substitution of $\mathcal{A}_{h_i}(\prec(h_i))$ for $S_{f(h_i)}$ in these formulae which gives the specified expression for the dual basis element in terms of Hall integrals. □

In this section we have provided formulae for the dual PBW basis elements alternative but equivalent to the ones to be found in the book [28].

5.5. A conjecture

Theorem 5.26 states that polynomials in *Hall integrals* freely generate the half shuffle algebra $(\mathcal{A}, \prec)$ as an associative and commutative algebra. A natural question is whether a similar structure theorem holds in the case where the half shuffle $\prec$ on Hall trees is replaced by the commutator area as basic operation. This question has been, and still remain, a conjecture well supported by calculation for the last decade.

Conjecture. *Any element of $\mathcal{A}$ can be written uniquely as a polynomial over Hall areas $\{\mathsf{area}(h)\}_{h\in H}$.*

Trying to solve this conjecture led us to consider an argument related to the well-known Lazard's elimination [28] to construct a canonical, but to our knowledge, new decomposition of the half shuffle algebra $(\mathcal{A}, \prec)$ as shuffle power series in the greatest letter c of the alphabet A with coefficients in a sub-algebra $\mathcal{X}$ freely generated by a new alphabet X with an infinite number of letters defined in terms of c and all other letters in A. This construction, that we refer to as elimination trick, allows us to provide, in the next section, a second proof relying on an induction argument of our structure theorem.

5.6. Another proof of the structure theorem

The following simple and concrete observation will be expanded in this section.

If $(\mathcal{A}, \prec)$ is the free half shuffle algebra over A, and $c \in A$, and X is the subset of $\mathcal{A}$ comprising $\frac{1}{k}{\prec}((ac^k))$, $a \in A \setminus c$, and Z is the space spanned by words that do not begin with c; then $(Z, \prec)$ is a half shuffle algebra generated by X in $\mathcal{A}$; moreover, Z is freely generated as a half shuffle algebra by X, and therefore canonically isomorphic as a half shuffle algebra to the free half shuffle algebra $(\mathcal{X}, \prec)$ over X. In characteristic zero, Z is the half shuffle sub-algebra of $\mathcal{A}$ spanned by the words that do not begin with c. It is complimentary to $\mathcal{A} \sqcup\!\sqcup\, c$ and we have

$$\begin{aligned} \mathcal{A} &= Z \oplus (\mathcal{A} \sqcup\!\sqcup\, c) \\ &= Z \oplus (Z \sqcup\!\sqcup\, c) \oplus (\mathcal{A} \sqcup\!\sqcup\, c \sqcup\!\sqcup\, c) \\ &= \dots \end{aligned} \tag{38}$$

and any element in $\mathcal{A}$ can be expressed canonically as a shuffle power series in c with coefficients in the half shuffle subalgebra Z. One can repeat this process by choosing a letter $d \in X$, and expanding every coefficient as a power series in d with coefficients in the half shuffle subalgebra generated by the elements $\{\frac{1}{k}{\prec}((ad^k)), a \in X \setminus d\}$. In what follows we will make this precise.

Definition 5.28. Let c be the greatest element of A with respect to an ancestral ordering $<$. Define the subset of trees

$$X = \{(ac^n), a \in A \setminus \{c\}, n \geq 0\} \subset \mathcal{M}_A. \tag{39}$$

With this choice of (infinite) alphabet, the following spaces and operators are automatically defined in the same way as their A counterparts:

- $\mathcal{M}_X$ the free magma;
- W_X the space of words in the alphabet X;
- $\mathcal{X}$ the vector space spanned by words in W_X;
- $\otimes_X$, $[\cdot,\cdot]_X$, $\prec_X$, area_X, $(\cdot,\cdot)_X$ the products and pairing on these spaces;
- $\mathcal{L}_X$ the free Lie sub-algebra of $(\mathcal{X}, \otimes_X)$;
- $\exp_X$, $\log_X$ the tensor series for the respective maps.

Remark 5.29. Note that the elements of W_X are words whose letters are particular words in A.

Theorem 5.30. *[28, Theorem 0.6] The Lie algebra $\mathcal{L}_A$ is the semi-direct product of $\mathcal{L}_X$ and $\mathbb{R}c$*

$$\mathcal{L}_A = \mathcal{L}_X \ltimes \mathbb{R}c. \tag{40}$$

As a result of Theorem 5.30, $\mathcal{L}_X$ is a Lie ideal and sub-algebra of co-dimension one in $\mathcal{L}_A$ and in particular $\mathcal{L}_A = \mathcal{L}_X \oplus \mathbb{R}c$.

Next we report an important lemma from [28] which provides a very simple relation between Hall sets in $\mathcal{M}_A$ and Hall sets in $\mathcal{M}_X$.

Lemma 5.31. *[28, Lemma 4.19 & Section 5.6.3] The unique homomorphism of magmas $\phi : \mathcal{M}_X \to \mathcal{M}_A$ that sends $x = (ac^n) \in X$ to $(ac^n) \in \mathcal{M}_A$ is an injection of magmas and its range is the free magma over X. Furthermore $\phi(H_X) = H \cap \phi(\mathcal{M}_X)$ is the Hall set in $\mathcal{M}_X$ associated with the ordering $<$, $H = \{c\} \cup \phi(H_X)$, and c is the greatest element of H.*

Remark 5.32. $\mathcal{M}_X$ is a sub-magma and inherits an ancestral ordering from $\mathcal{M}_A$. It follows that the image by ϕ of the Hall set H_X in $\mathcal{M}_X$ associated to the ordering $<$ is $H \cap \phi(\mathcal{M}_X)$ (Lemma 5.31).

When switching back and forth between the X- and A-spaces, the first objects one needs to have control over are letters from the two alphabets X and A. In the next lemma we express the images under the various operators discussed so far of letters in X, seen as trees in $\mathcal{M}_A$, in terms of words from W_A.

Lemma 5.33. *For any $x \in X$, the image $\phi(x)$ in $\mathcal{M}_A$ is of the form (ac^n) for some $a \in A$ and $n \geq 0$. The image of (ac^n) under the operators $[\,]$, $\prec$, area in $\mathcal{A}$, expressed in terms of words in W_A is given by*

$$[\phi(x)] = [(ac^n)] = \binom{n}{0} ac...c - \binom{n}{1} cac...c + ... + (-1)^n \binom{n}{n} c...ca \tag{41}$$

$$\prec(\phi(x)) = \prec((ac^n)) = n! ac...c \tag{42}$$

$$\mathsf{area}(\phi(x)) = \mathsf{area}((ac^n)) = n!(ac...c - cac...c) \tag{43}$$

where all the words are of length $n+1$ and contain exactly once the letter a.

The proof is left as an exercise to the reader.

The next lemma tells the relationship between integrals and areas on letters from X.

Lemma 5.34. *For any tree $(ac^n) \in \mathcal{M}_A$ one has*

$$\prec((ac^n)) = \frac{1}{n+1}\mathsf{area}((ac^n)) + \frac{n}{n+1} c \sqcup\!\sqcup \prec((ac^{n-1})). \tag{44}$$

Proof. From Lemma 5.33 we deduce the following identity

$$\mathsf{area}((ac^n)) + (c \sqcup\!\sqcup n \prec((ac^{n-1}))) = (n+1) \prec((ac^n))$$

which after rearranging yields equation (44). □

Lemma 5.35. *For any $(ac^n) \in \mathcal{M}_A$ one has*

$$\prec((ac^n)) = \frac{1}{n+1}\sum_{k=0}^{n} c^{\sqcup\!\sqcup k} \sqcup\!\sqcup \mathsf{area}((ac^{n-k})). \tag{45}$$

Proof. This follows immediately from Lemma 5.34 and an induction on n. □

Remark 5.36. Recall that the Lie bracket operator $[\cdot]$ is defined on $\mathcal{M}_A$ with values in $\mathcal{L}_A$. The restriction of $[\cdot]$ defined on $\mathcal{M}_A$ to $\mathcal{M}_X$ agrees with the natural definition of $[\cdot]$ on $\mathcal{M}_X$. It is also a simple exercise to prove that this compatibility between the restriction and the intrinsically defined operators holds for the tensor product and the Lie bracket.

Definition 5.37. We denote by $J_c : \mathcal{X} \to \mathcal{A}$ the unique $\prec$-homomorphism that, by freeness of $(\mathcal{X}, \prec_X)$ over X, extends to $\mathcal{X}$ the map

$$(ac^n) \mapsto \frac{1}{n}(\prec((ac^n))), \quad n > 0. \tag{46}$$

Denote by $(Z, \prec)$ the half shuffle subalgebra of $(\mathcal{A}, \prec)$ generated by the elements

$$\{J_c(x) : x \in X\}.$$

Next we prove that the algebra $(Z, \prec)$ is closed under $\prec$ and provide a characterisation of Z as the linear span of words in W_A that do not begin with the letter c.

Lemma 5.38. *Z is the span of words in W_A that do not begin with the letter c*

$$Z = Span\{w = a \prec v \in W_A \mid a \neq c, a \in A, v \in W_A\}.$$

In particular Z is closed under $\prec$.

Proof. Let Z' be the linear span in $\mathcal{A}$ of the $w \neq cv \in W_A$. It is immediate from the definitions of $\sqcup\!\sqcup$ and $\prec$ on words that Z' is closed under both operations. If $t = (t_1, t_2) \in \mathcal{M}_A$ and if, for $i = 1, 2$, $\prec(\phi(t_i))) \in Z'$ then $\prec(\phi(t)) = (\prec(\phi(t_1))) \prec (\prec(\phi(t_2)))$ is also in Z' because Z' is closed under $\prec$. Let $x \in X$, then by equation (42)

$$\prec(\phi(x)) = \prec((ac^n)) = n!ac...c \in Z'. \tag{47}$$

We may proceed recursively to see that every $\prec(t)$ contained in Z is also an element of Z'; since Z is generated by $\{J_c(x) : x \in X\}$ we conclude that $Z \subset Z'$. The unique decomposition of words into decreasing sequences of Hall words shows that the dimension of Z' and Z are equal, hence $Z = Z'$. □

Lemma 5.39. *The half shuffle algebra $\mathcal{A}$ has the following decomposition*

$$\mathcal{A} = Z \oplus (Z \sqcup\!\sqcup c) \oplus (Z \sqcup\!\sqcup c^{\sqcup\!\sqcup 2}) + ... \tag{48}$$

Proof. Consider any word $w \in W_A$ beginning with n number of c's.

$$w = c \prec (c \prec (... \prec (c \prec v)...))$$

where $v = av' \in Z$ is a word that doesn't begin with c, i.e. $a \in A$, $a \neq c$, $v' \in W_A$. If $n = 0$ then $w \in Z$. By induction on n

$$c \prec (c \prec (... \prec (c \prec v)...)) - \alpha_n c^{\sqcup\!\sqcup n} \sqcup\!\sqcup v = L$$

where $\alpha_n \in \mathbb{R}$ and L is a linear combination of words that begin with $k < n$ number of c's. Hence, by induction on the number of c's in front of the words, the word w can be written as a shuffle polynomial in c with coefficients in Z. □

Lemma 5.40. *J_c maps polynomials in Hall integrals $\prec_X(h)$, $h \in H_X$ to polynomials in Hall integrals $\prec(\phi(h))$.*

Proof. This follows immediately because J_c is a half shuffle (and so shuffle) homomorphism. □

We now repeat our structure theorem and provide an alternative proof based on the elimination trick discussed so far in this section.

Theorem 5.41. *The half shuffle algebra $(\mathcal{A}, \prec)$ is freely generated by polynomials in Hall integrals $\prec(h)$ for $h \in H$.*

Proof. We can assume by induction that the theorem holds for $\mathcal{X}$, i.e. that $(\mathcal{X}, \prec_X)$ is freely generated by polynomials in $\prec_X(h)$ for $h \in H_X$. By Lemma 5.40, Z is freely generate by polynomials in $\prec(h)$ with $h \in \phi(H_X)$. By Lemma 5.31, $H = \{c\} \cup \phi(H_X)$ and with the decomposition (48) we conclude that $\mathcal{A}$ is freely generated by polynomials in $\prec(h)$, $h \in H$. □

5.7. Scalable computations of path signatures

As mentioned in the introduction, instances of streamed information can be represented as a path $\gamma : [0, 1] \to V$ with values on some finite dimensional vector space $V \simeq \mathbb{R}^d$, such a path is faithfully represented, up to reparameterisation, by the signature $\mathcal{S}_\gamma \in (\mathcal{A}, \otimes)$. Furthermore, since the extended tensor algebra $(\mathcal{A}, \otimes)$ is the algebraic dual of the half shuffle algebra $(\mathcal{A}, \prec)$, it is automatic to see that the restriction of linear functionals on $\mathcal{A}$ to the range of the signature form a unital algebra of real-valued functions that separates signatures. Hence, by the Stone-Weierstrass theorem linear functionals acting on the signatures are dense in the space of continuous, real-valued functions on compact sets of unparameterised paths. Thus, non-linear regression on pathspace can be realised by linear regression on the terms of the signature. However, terms in the signature contain some redundancy, which represents a major scalability issue, particularly because the number of distinct and linearly independent iterated integrals grows exponentially in the truncation level. In this paper, and in particular in Theorems 5.26 and 5.41, we identified sets of Hall integrals that can be used to compute any term in the signature with a minimal amount of computations.

To illustrate this we consider a simple example. Let $d = 3$ and let us identify the 3-dimensional vector space V as the space spanned by an alphabet of three letters $A = \{1, 2, 3\}$. Let $\omega = \mathbf{233212222111}$; note that $|\omega| = 12$. Then, computing the coefficient $(S_\omega, \mathcal{S}_\gamma)$ in the signature using existing software [19,22,27] (based on the *Chen's relation*) involve evaluating the level-12 truncated tensor exponential of increments $\exp^{(12)}(\gamma_t - \gamma_s)$. This operation has space and time complexities of $\mathcal{O}(3^{12})$.

Instead, considering for example the Lyndon basis, one can precompute the factorisation of ω into decreasing product of Lyndon words and find

$$\omega = f(\mathbf{1})^{\otimes 3} \otimes f(((((\mathbf{1}, \mathbf{2}), \mathbf{2}), \mathbf{2}), \mathbf{2})) \otimes f(\mathbf{2}) \otimes f(((\mathbf{2}, \mathbf{3}), \mathbf{3}))$$

Therefore, by Theorem 5.26 one has

$$S_\omega = \prec(\mathbf{1})^{\sqcup\!\sqcup 3} \sqcup\!\sqcup \frac{1}{4!} \prec(((((\mathbf{1}, \mathbf{2}), \mathbf{2}), \mathbf{2}), \mathbf{2})) \sqcup\!\sqcup \prec(\mathbf{2}) \sqcup\!\sqcup \frac{1}{2!} \prec(((\mathbf{2}, \mathbf{3}), \mathbf{3})).$$

Using the interplay between algebraic operations $\prec$ and $\sqcup\!\sqcup$ and the rules of calculus on paths outlined in Section 3 we obtain

$$(S_\omega, \mathcal{S}_\gamma) = \frac{1}{48} \alpha_1 \alpha_2 \alpha_3 \alpha_4$$

where

$$\begin{aligned}
\alpha_1 &= \int_0^1 d\gamma_t^{(1)} \\
\alpha_2 &= \int_0^1 \left(\int_0^v \left(\int_0^u \left(\int_0^t \gamma_s^{(1)} d\gamma_s^{(2)} \right) d\gamma_t^{(2)} \right) d\gamma_u^{(2)} \right) d\gamma_v^{(2)} \\
\alpha_3 &= \int_0^1 d\gamma_t^{(2)} \\
\alpha_4 &= \int_0^1 \left(\int_0^t \gamma_s^{(2)} d\gamma_s^{(3)} \right) d\gamma_t^{(3)}.
\end{aligned}$$

6. Conclusion

In this paper, we identified the free Zinbiel algebra introduced by [31] with an algebra of real-valued functions on paths. We provided two, to our knowledge, new basic identities in arity 3 involving its symmetrisation $\sqcup\!\sqcup$ and its anti-symmetrisation area. We showed that these are sufficient to recover the Zinbiel and Tortkara identities introduced by Dzhumadil'daev [11]. We then used these identities to provide a direct proof of the main result in [8] stating that polynomials in iterated areas generate the free Zinbiel algebra [32]. Subsequently, we introduced minimal sets of Hall integrals and showed, with two different proof techniques, that polynomial functions on these Hall integrals freely generate the half shuffle algebra. This result can be interpreted as a structure theorem for streamed information, allowing to split real valued functions on streamed data into two parts: a first that extracts and packages the streamed information into Hall integrals, and a second that evaluates a polynomial in these without further reference to the original stream.

Data availability

No data was used for the research described in the article.

Acknowledgments

We deeply thank Prof. Pavel Kolesnikov and Prof. Frédéric Patras for the helpful discussions and suggestions. Terry Lyons and C. Salvi's contributions to this work are supported by the EPSRC program grant DataSig [grant number EP/S026347/1]. Terry Lyons' contribution was also supported by project partners: in part by The Alan Turing Institute under the EPSRC grant EP/N510129/1, in part by The Alan Turing Institute's Data Centric Engineering Programme under the Lloyd's Register Foundation grant G0095, in part by The Alan Turing Institute's Defence and Security Programme, funded by the UK Government, and in part by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA).

References

[1] I.P. Arribas, C. Salvi, L. Szpruch, Sig-sdes model for quantitative finance, in: ACM International Conference on AI in Finance, 2020.

[2] H. Boedihardjo, X. Geng, T. Lyons, D. Yang, The signature of a rough path: uniqueness, Adv. Math. 293 (2016) 720–737.

[3] N. Bourbaki, Lie Groups and Lie Algebras, chapters 2-3, Springer Science & Business Media, 2008.

[4] T. Cass, W.F. Turner, Topologies on unparameterised path space, arXiv preprint arXiv:2206.11153, 2022.

[5] K.-T. Chen, Integration of paths, geometric invariants and a generalized Baker-Hausdorff formula, Ann. Math. (1957) 163–178.

[6] N.M. Cirone, M. Lemercier, C. Salvi, Neural signature kernels as infinite-width-depth-limits of controlled resnets, arXiv preprint arXiv:2303.17671, 2023.

[7] T. Cochrane, P. Foster, V. Chhabra, M. Lemercier, T. Lyons, C. Salvi, Sk-tree: a systematic malware detection algorithm on streaming trees via the signature kernel, in: 2021 IEEE International Conference on Cyber Security and Resilience (CSR), IEEE, 2021, pp. 35–40.

[8] J. Diehl, T. Lyons, R. Preiß, J. Reizenstein, Areas of areas generate the shuffle algebra, arXiv preprint arXiv:2002.02338, 2020.

[9] J. Diehl, J. Reizenstein, Invariants of multidimensional time series based on their iterated-integral signature, Acta Appl. Math. 164 (1) (2019) 83–122.

[10] A. Dzhumadil'daev, N. Ismailov, F. Mashurov, On the speciality of tortkara algebras, J. Algebra 540 (2019) 1–19.

[11] A. Dzhumadil'daev, Zinbiel algebras under q-commutators, J. Math. Sci. 144 (2) (2007) 3909–3925.

[12] K. Ebrahimi-Fard, F. Patras, Cumulants, free cumulants and half-shuffles, Proc. R. Soc. A 471 (2176) (2015) 20140843.

[13] A. Fermanian, T. Lyons, J. Morrill, C. Salvi, New directions in the applications of rough path theory, IEEE BITS Inform. Theory Mag. (2023).

[14] E. Gehrig, M. Kawski, A Hopf-algebraic formula for compositions of noncommuting flows, in: 2008 47th IEEE Conference on Decision and Control, IEEE, 2008, pp. 1569–1574.

[15] B. Hambly, T. Lyons, Uniqueness for the signature of a path of bounded variation and the reduced path group, Ann. Math. (2010) 109–167.

[16] B. Horvath, M. Lemercier, C. Liu, T. Lyons, C. Salvi, Optimal stopping via distribution regression: a higher rank signature approach, arXiv preprint arXiv:2304.01479, 2023.

[17] M. Kawski, Chronological algebras: combinatorics and control, in: Geometric Control Theory, Moscow, 1998, in: Itogi Nauki Tekh. Ser. Sovrem. Mat. Prilozh. Temat. Obz., vol. 64, Vseross. Inst. Nauchn. i Tekhn. Inform. (VINITI), Moscow, 1999, pp. 144–178 (in Russian).

[18] P. Kidger, P. Bonnier, I. Perez Arribas, C. Salvi, T. Lyons, Deep signature transforms, Adv. Neural Inf. Process. Syst. 32 (2019).

[19] P. Kidger, T. Lyons, Signatory: differentiable computations of the signature and logsignature transforms, on both CPU and GPU, arXiv:2001.00706, 2020.

[20] M. Lemercier, C. Salvi, T. Cass, E.V. Bonilla, T. Damoulas, T. Lyons, Siggpde: scaling sparse Gaussian processes on sequential data, in: International Conference on Machine Learning, PMLR, 2021.

[21] M. Lemercier, C. Salvi, T. Damoulas, E. Bonilla, T. Lyons, Distribution regression for sequential data, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 3754–3762.

[22] T. Lyons, et al., Coropa computational rough paths (software library), 2010.

[23] T. Lyons, M. Caruana, T. Lévy, Differential equations driven by rough paths, in: Ecole d'été de Probabilités de Saint-Flour, vol. XXXIV, 2004, pp. 1–93.

[24] T.J. Lyons, Differential equations driven by rough signals, Rev. Mat. Iberoam. 14 (2) (1998) 215–310.

[25] J. Morrill, C. Salvi, P. Kidger, J. Foster, Neural rough differential equations for long time series, in: International Conference on Machine Learning, PMLR, 2021, pp. 7829–7838.

[26] R. Ree, Lie elements and an algebra associated with shuffles, Ann. Math. (1958) 210–220.

[27] J. Reizenstein, B. Graham, The iisignature library: efficient calculation of iterated-integral signatures and log signatures, arXiv preprint, arXiv:1802.08252, 2018.

[28] C. Reutenauer, Free Lie Algebras, London Mathematical Society Monographs. Oxford Science Publications, The Clarendon Press, Oxford University Press, 1993.

[29] C. Salvi, T. Cass, J. Foster, T. Lyons, W. Yang, The signature kernel is the solution of a Goursat pde, SIAM J. Math. Data Sci. 3 (3) (2021) 873–899.

[30] C. Salvi, M. Lemercier, C. Liu, B. Horvath, T. Damoulas, T. Lyons, Higher order kernel mean embeddings to capture filtrations of stochastic processes, Adv. Neural Inf. Process. Syst. 34 (2021) 16635–16647.

[31] M.P. Schützenberger, Sur une propriété combinatoire des algebres de Lie libres pouvant être utilisée dans un probleme de mathématiques appliquées, in: Séminaire Dubreil, Algèbre Théorie Nr. 12 (1) (1958) 1–23.

[32] H. Sussmann, A product expansion for the Chen series, in: C.I. Byrnes, A. Lindquist (Eds.), Theory and Applications of Nonlinear Control Systems, North-Holland, 1986, pp. 323–335.